LOCALIZING THE EIGENVALUES OF MATRIX-VALUED FUNCTIONS: ANALYSIS AND APPLICATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Amanda Hood

January 2017

© 2017 Amanda Hood
ALL RIGHTS RESERVED

LOCALIZING THE EIGENVALUES OF MATRIX-VALUED FUNCTIONS: ANALYSIS AND APPLICATIONS

Amanda Hood, Ph.D.

Cornell University 2017

The standard eigenvalue problem of finding the zeros of $\det(zI-A)$ is ubiquitous, and comes from studying solutions to $x' = Ax$ as well as myriad other sources. Similarly, if $x'(t) = Ax(t) + Bx(t-1)$ is some delay-differential equation (arising, say, from modeling the spread of a disease, or from population growth models), then stability is determined by computing the roots of $\det(zI - A - Be^{-z})$. Models of physical processes where energy slowly leaks away to infinity lead to similar problems. These physical systems are typically modeled in terms of differential equations which are then discretized using e.g. collocation or finite elements. For example, such a discretization gives a correspondence between the quantum scattering resonances associated to a quantum corral and the zeros of $\det(A-zB+C(z))$, where $A, B, C(z) \in \mathbb{C}^{n\times n}$ and the highly nonlinear entries of $C(z)$ involve square roots and ratios of Bessel functions. In each case, we are led to so-called *nonlinear eigenvalue problems* of the form $T(\lambda)v = 0$, $v \neq 0$, where $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$ is a matrix-valued function, and λ is called an eigenvalue of T.

The first contribution of this thesis is theorems for localizing eigenvalues of general matrix-valued functions, effectively reducing the region in which eigenvalues of T are known to lie from all of Ω down to a smaller space, and deducing eigenvalue counts within regions that meet certain conditions. These theorems are derived by working with the diagonal entries or diagonal blocks of T, such as our generalization of Gershgorin's theorem, or by considering nonlinear gen-

eralizations of pseudospectra. Localization and counting results allow better initial guesses or shifts for iterative algorithms, guide the selection of an appropriate closed contour for contour integral-based algorithms, and facilitate error analysis in cases where eigenvalues can be confined to tiny regions.

The second contribution of this thesis is to exploit these results in several contexts. To start with, a variety of strategies for getting the most out of our main localization theorems will be presented and applied to several test problems. Then we foray into the analysis of higher-order and delay differential equations, using our localization results to help bound *asymptotic* growth of solutions, and using our generalization of the notion of pseudospectra to concretely bound *transient* growth both above and below; a model for a semiconductor laser with phase-conjugate feedback acts as the central example. The last application we will treat is about the resonances for electrons trapped in circular quantum corrals, microscopic structures built by placing atoms in a circle on a metal surface. We provide a framework for comparing various elastic-scattering models, and use it to bound the error between resonances computed from the naïve particle-in-a-box model and resonances computed from a model that takes quantum tunneling into account.

BIOGRAPHICAL SKETCH

Amanda Hood studied mathematics at Rutgers University, graduating summa cum laude in 2009 with a B.A. and high honors from the Department of Mathematics. While there, she was introduced to numerical linear algebra by way of research on gravitational lens modeling. After entering the Center for Applied Mathematics at Cornell University, she worked with her advisor David Bindel in the area of numerical linear algebra with a focus on nonlinear eigenvalue problems. She received her M.S. in applied mathematics in May 2013. At roughly the same time, she received an Outstanding Teaching Assistant award and was co-author on her first paper, which was doubly honored in 2015 with the SIAG/Linear Algebra Prize and selection as a SIGEST paper in the journal SIAM Review.

To all the teachers who challenged and encouraged me.

ACKNOWLEDGEMENTS

Work in this thesis was supported in part by the Sloan Foundation, by the National Science Foundation under Grant No. DMS-1620038, and by Cornell University through the Sage Fellowship. In addition to these organizations, I thank my advisor, David Bindel, for introducing me to many new ideas, and Eric and my family for their support.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vi
List of Tables . . . viii
List of Figures . . . ix

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Some vocabulary . . . 7
1.3 A brief history of localization . . . 12
1.4 Contributions and outline . . . 20

2 Localization theorems for general matrix-valued functions **23**
2.1 Gershgorin theorems . . . 23
2.1.1 Inspiration . . . 23
2.1.2 Useful tools . . . 25
2.1.3 Extensions for matrix-valued functions . . . 29
2.2 Bauer-Fike theorems . . . 34
2.3 Pseudospectral localization theorems . . . 38
2.3.1 Recommendations . . . 41

3 Gallery of examples **46**
3.1 Naïve application of nonlinear Gershgorin theorem . . . 46
3.1.1 Single delay PDDE I . . . 46
3.1.2 Fiber . . . 48
3.1.3 Planar waveguide . . . 50
3.1.4 Butterfly I . . . 51
3.2 Diagonalizing the dominant term . . . 52
3.2.1 Single delay PDDE II . . . 52
3.2.2 CD player . . . 55
3.2.3 Butterfly II . . . 57
3.2.4 Hospital . . . 59
3.2.5 HIV I . . . 61
3.3 Using the Cholesky decomposition . . . 64
3.3.1 Loaded string . . . 64
3.3.2 Butterfly III . . . 66
3.4 Dropping terms . . . 67
3.4.1 Hadeler . . . 67
3.5 Using block structure . . . 72
3.5.1 Delta potentials . . . 72
3.5.2 Butterfly IV . . . 79

3.6 Using approximations . . . 80
3.6.1 Gun . . . 81
3.6.2 HIV II . . . 83

4 Transient dynamics **86**
4.1 Introduction . . . 86
4.2 Preliminaries . . . 88
4.3 Upper bounds for higher-order ODEs . . . 91
4.4 Upper bounds for delay differential equations . . . 97
4.5 Lower bounds . . . 105
4.6 Conclusion . . . 107

5 Scattering resonances and quantum corrals **108**
5.1 Introduction . . . 108
5.2 Background . . . 110
5.3 Discretization . . . 115
5.4 Error analysis . . . 119
5.4.1 Error due to discretization . . . 119
5.4.2 Error due to rational approximation . . . 121
5.4.3 Truncation of the sequence of 1D problems . . . 123
5.5 Examples . . . 125
5.5.1 The particle in a box . . . 125
5.5.2 Rational approximation to the DtN map . . . 130
5.6 Conclusion . . . 134

6 Conclusion **135**

A Delta potentials **136**

B Change of variable making gun polynomial **141**

C The Dirichlet-to-Neumann map **144**

D Derivation of realistic quantum tunneling model parameters **153**

Bibliography **155**

LIST OF TABLES

3.1 Error table for single delay PDDE I. 54
3.2 Error table for the `hospital` problem. 62
3.3 Error table for the `hadeler` problem. 71

5.1 Some eigenvalues of $T_n^{(\mathrm{Dir})}$ for various n and their first order corrections according to Proposition 5.3. The residual at k is defined to be $\sigma_{\min}\left(T^{(\mathrm{DtN})}(k)\right)$. 129

LIST OF FIGURES

1.1 Eigenvalues of `gun` problem from [Lia07, Tables 6.3, 6.5]. 5
1.2 Localization regions for `gun` problem. 5
1.3 Comparison with analysis from [Eff13, §5.1]. 6

2.1 Inclusion regions for hypothesis-illustrating examples. 32

3.1 Naïve inclusion regions for a single delay PDDE. 48
3.2 Naïve inclusion region for `fiber` problem and graph of $s(z)$. . . 48
3.3 Inclusion regions for `planar_waveguide` problem. 50
3.4 Spectrum and naïve inclusion regions for `butterfly` problem. . 52
3.5 Diagonalization-based inclusion regions and eigenvalues for single delay PDDE example. 54
3.6 Inclusion regions for `cd_player` problem. 56
3.7 Comparison with matrix polynomial inclusion regions for the `cd_player` problem. 57
3.8 Localization region from simultaneous diagonalization of z^4, z^2, and constant terms in `butterfly` problem. 59
3.9 Eigenvalues and inclusion regions for the `hospital` problem. . 60
3.10 Localization region and envelope curves for HIV problem. 63
3.11 Localization regions and approximate eigenvalues for the `loaded_string` problem. 65
3.12 Localization regions for `butterfly` problem obtained with Cholesky decomposition of A_4 and diagonalization of another term. 67
3.13 Inclusion regions for the `hadeler` problem using different splittings. 70
3.14 Using Beyn's Integral Algorithm 2 to compute eigenvalues for the `hadeler` problem. 72
3.15 Square roots of resonances for $V(x) = \delta(x) + \delta(x-1)$. 73
3.16 Inclusion regions and pseudospectra for some problems with 2 delta potentials of equal strength. 77
3.17 Inclusion regions and pseudospectra for problems with 3 and 10 delta potentials of differing strengths. 78
3.18 Structure of matrices in `butterfly` problem. 79
3.19 Localization region for `butterfly` problem obtained by applying the nonlinear block Gershgorin theorem. 80
3.20 Exact and approximate eigenvalues and localization regions for `gun` problem. 82
3.21 Deriving pseudospectral inclusion for the HIV problem. 84

4.1 Discretization-based upper bounds for laser example. 96
4.2 Upper bounds for discretized partial delay differential equation example and the contours from which they are derived. 103

4.3 Upper bound for laser example and the contour from which it is derived. 105
4.4 Lower bounds for discretized partial delay differential equation example (left) and laser example (right). 107

5.1 Eigenvalues of matrix-valued function $T_n^{(\mathrm{Dir})}(k)$ for index $n = 0, 1, 2, 3$, constructed using 40 and 10 points on $[0, R_1]$ and $[R_1, R]$, resp., for $R_1 = 0.95$, $R = 1$, with plot of corresponding Bessel function $J_n(kR)$. Resonance estimates are $E = k^2$ as defined by (5.32). . 126
5.2 Illustration of two potentials with mutually consistent meshes. . 127
5.3 Pseudospectra and error norms used in comparing the particle in a box model to a quantum tunneling model. 127
5.4 Plot used to deduce inclusion regions and counts for resonances of quantum tunneling model. 128
5.5 Pseudospectra (left) for realistic quantum tunneling model and error (right) between that model and the particle in a box. 131
5.6 The ellipse used to define the rational approximation and contour plot of $\log_{10}|f(z) - f_n(\sqrt{z}, R)|$ for $n = 0$ (left). The square root of the ellipse and contour plot of $\log_{10}\|T_0^{(\mathrm{DtN})}(k) - T_0^{(\mathrm{rat})}(k)\|$ (right), where $T_0^{(\mathrm{rat})}(k)$ is defined according to (5.14). 131
5.7 Level curves of $\log_{10}\sigma_{\min}T_n^{(\mathrm{DtN})}(k)$ (green) and $\log_{10}\|T_n^{(\mathrm{DtN})}(k) - T_n^{(\mathrm{rat})}(k)\|$ (red) at -1 (left) and -1.3 (right). Compare with Figure 5.6 (right). 132
5.8 Rectangular region of interest, contours of $\log_{10}\|T_n^{(\mathrm{DtN})}(\sqrt{E}) - T_n^{(\mathrm{rat})}(\sqrt{E})\|_2$, and eigenvalues of $T_n^{(\mathrm{rat})}(\sqrt{E})$, all for $n = 0$ (left). The contours $\log_{10}\|T_n^{(\mathrm{DtN})}(\sqrt{E}) - T_n^{(\mathrm{rat})}(\sqrt{E})\|_2 = -1.8$ (red) and $\log_{10}\left(\sigma_{\min}\left(T_n^{(\mathrm{DtN})}(\sqrt{E})\right)\right) = -1.8$ (green), and the eigenvalues of $T_n^{(\mathrm{rat})}(\sqrt{E})$ in the rectangle of interest (right). 133

CHAPTER 1

INTRODUCTION

1.1 Motivation

The standard eigenvalue problem of finding the zeros of $\det(zI - A)$ is ubiquitous, and comes from studying solutions to $x' = Ax$ as well as myriad other sources. Somewhat less well-known is the analogous problem arising in the study of delay differential equations, which are used to model processes such as the spread of a disease with an incubation period, or population growth models where gestation period is taken into account. Specifically, it is the zeros of $\det(zI - A - Be^{-z})$ that determine asymptotic growth and decay for solutions of the delay differential equation $x'(t) = Ax(t) + Bx(t-1)$ [MN07b, Prop. 1.12].

Models of physical systems where energy is lost lead to similar problems. For a second order differential equation with a damping term, such as $M\ddot{x} + C\dot{x} + Kx = f$, solution behavior is determined by the zeros of the determinant of the quadratic expression $Mz^2 + Cz + K$ [FT01]. More complicated expressions arise from systems where energy slowly leaks away to infinity via a radiation process. These physical systems are typically modeled in terms of differential equations which are then discretized using collocation or finite elements. For example, such a discretization gives a correspondence between the scattering resonances associated to a quantum corral and the zeros of $\det(A - zB + C(z))$, where $A, B, C(z) \in \mathbb{C}^{n\times n}$ and the highly nonlinear entries of $C(z)$ involve square roots and ratios of Bessel functions (Chapter 5).

The clear commonality among these cases is the characterization in terms of

determinants of matrices. More precisely, $zI - A$, $zI - A - Be^{-z}$, $Mz^2 + Cz + K$, and $A - zB + C(z)$ are all *matrix-valued functions*, which are mappings $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{m \times n}$ with $m = n$ for the examples above. In keeping with the definition of an eigenvalue of a matrix A as a number λ such that $\det(\lambda I - A) = 0$, an eigenvalue of a matrix-valued function T is defined to be a number λ such that $\det T(\lambda) = 0$. If $T(\lambda)v = 0$ with $v \neq 0$, then v is called an eigenvector for the eigenvalue λ. To distinguish it from the ordinary eigenvalue problem, finding eigenvalues of a given matrix-valued function T is often called a *nonlinear eigenvalue problem*.[1] Nonlinear eigenvalue problems arise in diverse fields and in innumerable forms. We have already mentioned a few, and we will analyze several examples in Chapters 3, 4, and 5. Among the problems we take from the NLEVP collection [BHM+13] are applications in fiber optic design (`fiber`, [Kau06]), particle accelerator design (`gun`, [Lia07]), control systems (`cd_player`, [DSB92]), modelling buildings (`hospital`, [CD02, Building Model]), and the computation of eigenvibrations on a string with an elastically attached load (`loaded_string`, [Sol06]). We also analyze an example from biology [CR00], stability for a delay differential equation [Eff13], [Jar08], two examples from quantum scattering, and an example from laser physics [GW06]. Excellent overviews of nonlinear eigenvalue problems and further examples are found in [FT01] and [MV04].

Though similar in definition, the process of computing eigenvalues for a matrix-valued function is significantly more treacherous than the computation of eigenvalues of a matrix. One obstacle is that eigenvectors associated to distinct eigenvalues of a matrix-valued function need not be linearly indepen-

[1] More properly, this is one type of nonlinear eigenvalue problem. The same phrase has been adopted for more general problems, including those in the continuous setting [AR10, FS68]. Since our results do not generalize to these cases, the term "nonlinear eigenvalue problem" is generally avoided in this thesis.

dent. Thus, the consensus seems to be that when one is interested in computing several eigenvalues of a matrix-valued function with an iterative method, one should not work with sets of eigenvalues and eigenvectors, but rather invariant pairs. See [Eff13], [Kre09], [BEK11], and [JMM14] for instances of iterative methods for computing invariant pairs. Also see [Vos13] for a self-contained overview of these and other iterative methods for nonlinear eigenvalue problems, and [VBJM16] and [BMM13] for more recent developments. In every case, an iterative method requires the user to provide either a number of initial guesses or a shift for which the nearest eigenvalues are of interest. There are contour integration-based methods that avoid the eigenvector issue as well, in [AST+09] and [Bey12], and more recently the variations in [XMZZ16] and [VBK16]. For these, the user must provide a simple closed curve enclosing eigenvalues of interest. Clearly, then, having some idea ahead of time as to where a matrix-valued function's eigenvalues are or are not is quite helpful in using the state of the art algorithms for computing them.

Another difficulty comes from the fact that the eigenvalues of a matrix-valued function T can be infinite in number. Therefore, without some information about their locations and counts, it is quite possible to miss eigenvalues of interest, perhaps by stopping an iterative algorithm too early, using the wrong initial guesses, or by poorly choosing a contour or parameters for integration-based methods so that eigenvalues of interest are excluded or computed inaccurately. In addition, sometimes we are not interested so much in the exact values of eigenvalues, but would rather have a quick test as to whether they are e.g. all in the left half-plane. For instance, we may want to check this for a matrix-valued function like $T(z) = zI - A - Be^{-z}$ for the purpose of a stability analysis [MN07b, Prop. 1.6]. Obviously this can't be done by computing

the (generically) infinitely many eigenvalues of T, and we must resort to *localization*, i.e., proofs about where eigenvalues could conceivably be (inclusion regions) and where they cannot exist (exclusion regions).

Before concluding this section, we briefly illustrate the usefulness of methods employed in this thesis on two problems from the literature. We use analysis performed by other authors for comparison.

The first of these examples is the `gun` problem from [BHM+13]. The original source [Lia07] contains an experiment where the 10 eigenvalues of $F(\lambda) = K - \lambda M + i\sqrt{\lambda}W_1 + i\sqrt{\lambda - 108.8774^2}W_2$ with $\sqrt{\lambda}$ nearest the shift 146.71 are to be computed, the square root being the principal branch. Putting $z = \sqrt{\lambda}$ and $T(z) = F(\lambda)$ for ease of discussion, Tables 6.3 and 6.5 in [Lia07] each contain a set of initial guesses and computed eigenvalues of T. By the author's own assessment [Lia07, p. 61, 2.], Table 6.3 is missing a few of the desired eigenvalues of T and instead includes some which are far from the shift 146.71. Furthermore, Table 6.3 contains a certain eigenvalue of T which is one of the ten closest to the shift, but this eigenvalue is missing from Table 6.5 (although it turns out this particular eigenvalue fails to meet some other conditions, rendering it undesirable anyway). Figure 1.1 shows the initial guesses ($\circ$) and computed eigenvalues ($*$) from Table 6.3 (left) and Table 6.5 (right). These two plots are very different, and do not in themselves lend confidence to the idea that all the eigenvalues of interest have been found via these two computations. Contrast this with the analysis we perform in Chapter 3, where we prove that all eigenvalues of T in $[120, 350] \times [-20, 50] \subset \mathbb{C}$ are contained in the thick blue contours shown in Figure 1.2 and easily count how many eigenvalues of T are in each. Not only does this information potentially allow us to use state of the art algorithms more

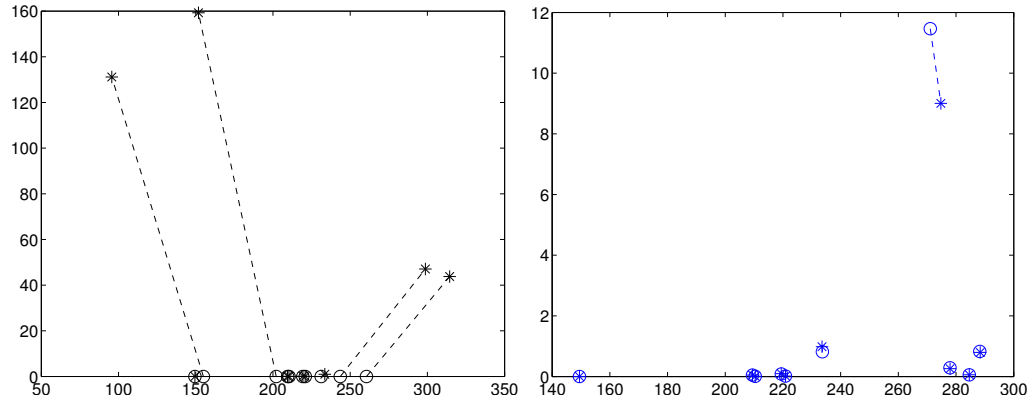

Figure 1.1: Eigenvalues of `gun` problem from [Lia07, Tables 6.3, 6.5].

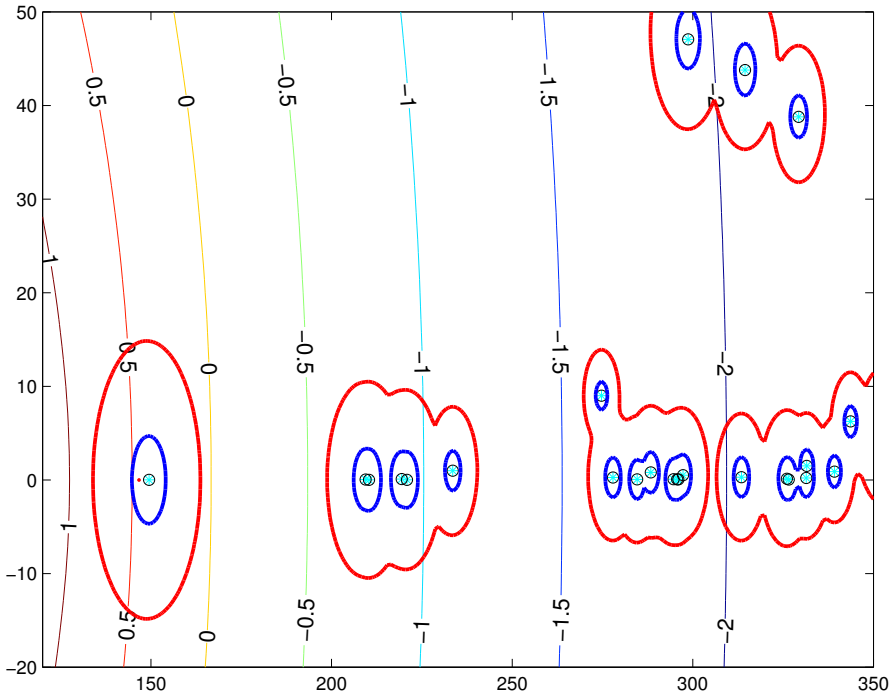

Figure 1.2: Localization regions for `gun` problem.

wisely in computing the eigenvalues of T, but it also concretely justifies that Table 6.5 in [Lia07] contains all 10 eigenvalues of T in $[120, 350] \times [-20, 50]$ that are closest to the shift 146.71.

The second example comes from an asymptotic stability analysis of a partial differential equation, performed in [Eff13, §5.1].[2] The authors compute the eight

[2] This problem is analyzed in Chapter 3 where it is referred to as the "single delay PDDE".

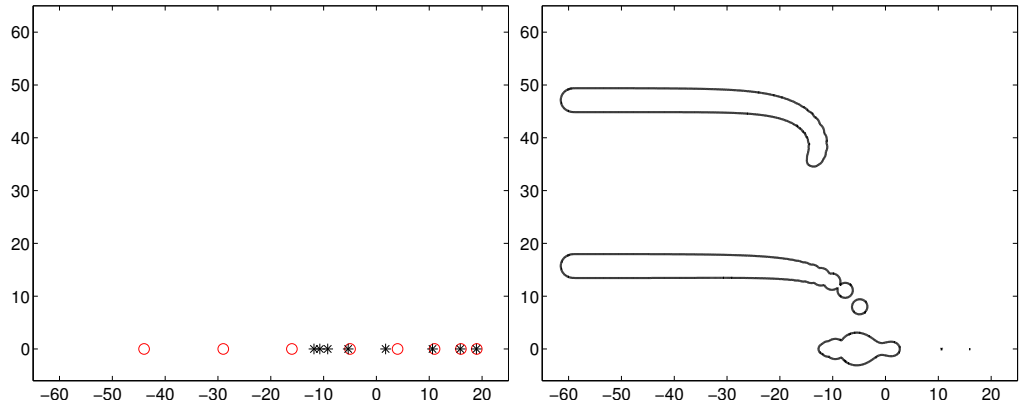

Figure 1.3: Comparison with analysis from [Eff13, §5.1].

largest real eigenvalues of a matrix-valued function $T(z) = zI - A_0 - A_1 e^{-\tau z}$ using naïve initial guesses. This is presumably on purpose, in order to demonstrate that the robust method presented in [Eff13] can give good results even with questionable input. See Figure 1.3 (left) for the initial guesses used (red circles) and the eigenvalues that were computed (black stars). Compare this with the inclusion regions we compute in Chapter 3, shown in Figure 1.3 (right), which must contain all eigenvalues of T within the pictured rectangular subset of $\mathbb{C}$. The inclusion region component around the eigenvalue ≈ 19 is so tiny it's not visible, and the components around eigenvalues at ≈ 11 and ≈ 16 are barely so. In addition, our analysis also shows exactly how many eigenvalues of T are in each component: in particular, the tiny components each contain one, and the origin-containing component contains 5. Therefore plotting the inclusion regions in Figure 1.3 (right) would have made it clear from the outset that the three leftmost initial guesses in Figure 1.3 (left) were inappropriate, and would have allowed initial guesses to be chosen in a more methodical way.

We also can easily show that the inclusion region components at $\approx -4.6 + 8.1i$ (the inclusion regions for this problem are symmetric about the real axis, though the mirror images are not shown) each contain one eigenvalue of T. From this it

is obvious that the eigenvalues computed in [Eff13, §5.1] are slightly misleading if one hopes to use them in stability analysis, because the eigenvalues of T associated to fastest asymptotic growth (or slowest asymptotic decay) are the ones with *largest real part* [MN07b, Prop. 1.12]. In particular, the eigenvalues in the components near $\approx -4.6 \pm 8.1i$ are more important for asymptotic analysis than e.g. the eigenvalue at ≈ -11.8. Thus, the localization region also could have been used to qualitatively understand whether more than just the real eigenvalues of T should be computed in a stability analysis.

1.2 Some vocabulary

In this section we present the terminology that will be used throughout this thesis. Formal definitions that are relevant to theorems used in the remainder of this work will be numbered. Informal definitions will also be scattered throughout this section and emphasized with italic text. When our terms differ from those used in other sources, it will be pointed out.

Our central object of study is the *nonlinear eigenvalue problem* of computing $\lambda \in \mathbb{C}$ such that $T(\lambda)v = 0$ for some nonzero vector v, where $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$ is analytic and *regular*. For T to be regular, $\det T(z)$ cannot be identically zero on any connected component of the domain Ω. The function T may be referred to as a *matrix function*, as in [Vos13], and sometimes as a λ-matrix (see [FT01, §3.3] and references therein). Here we prefer the term *matrix-valued function*.

Definition 1.1 (Matrix-valued function)**.** *A matrix-valued function with domain $\Omega \subset \mathbb{C}$ is a map $T : \Omega \to \mathbb{C}^{m\times n}$. If each entry function $T_{ij} : \Omega \to \mathbb{C}$ is analytic, then T is called analytic. If $m = n$, then T is called square. If $\det(T(z))$ is not identically zero on*

any connected component of Ω, then T is called regular.

In this thesis, the reader may assume unless otherwise specified that all matrix-valued functions under discussion are square, analytic, and regular.

Notice that if A is an $n \times n$ matrix then $z \mapsto zI - A$ is a matrix-valued function. By analogy with the usual notion of eigenvalue, we define (finite) eigenvalues of matrix-valued functions as follows.[3] We confine our attention to finite eigenvalues unless otherwise specified.

Definition 1.2 (Eigenvalue, eigenvector, spectrum)**.** *If $\lambda \in \mathbb{C}$ satisfies $\det T(\lambda) = 0$, then λ is called an eigenvalue of T. Equivalently, if there is a nonzero vector v such that $T(\lambda)v = 0$, then λ is an eigenvalue and v is then called an eigenvector; λ, v will be referred to as an eigenpair. The set of all eigenvalues of T is called its spectrum and denoted by $\Lambda(T)$.*

A matrix-valued function may have any number of eigenvalues, and eigenvectors associated to distinct eigenvalues need not be linearly independent. As examples, consider

$$T_1(z) = \begin{bmatrix} e^z & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad T_2(z) = \begin{bmatrix} (z-1)(z-2)(z-3) & 0 \\ 0 & 1 \end{bmatrix}.$$

T_1 has no eigenvalues and T_2 has three, and each eigenvalue of T_2 has $[1, 0]^T$ as an eigenvector. Despite these complications, we still have a notion of algebraic multiplicity.

Definition 1.3 (Algebraic multiplicity)**.** *We say an eigenvalue λ of $T(z)$ has algebraic multiplicity m if the Taylor expansion of $f(z) = \det(T(z))$ at λ is $f(z) = \sum_{j=m}^{\infty} a_j(z-\lambda)^j$*

[3] Although it is not common, some (see [Jar08, Ch. 4]) have found it convenient to define an eigenvalue of a matrix-valued function G as a complex number λ such that $G(\lambda)v = \lambda v$ for some nonzero vector v.

with $a_m \neq 0$, i.e.,

$$\left.\frac{d^j}{dz^j}\right|_\lambda \det(T(z)) = 0, \qquad j = 0, 1, ..., m-1, \tag{1.1}$$

$$\left.\frac{d^m}{dz^m}\right|_\lambda \det(T(z)) \neq 0. \tag{1.2}$$

We will call the problem of finding the eigenvalues of a matrix A, i.e. finding the eigenvalues of $z \mapsto zI - A$, the *ordinary eigenvalue problem* or the *standard eigenvalue problem*. Other common types of eigenvalue problems are listed below, along with the form of the corresponding matrix-valued function:

- **generalized**: $z \mapsto zB - A$, $A, B \in \mathbb{C}^{n\times n}$
- **quadratic**: $Q(z) = A_0 + A_1 z + A_2 z^2$, $A_j \in \mathbb{C}^{n\times n}$
- **polynomial**: $P(z) = \sum_{j=0}^{n} A_j z^j$, $A_j \in \mathbb{C}^{n\times n}$
- **rational**: $R(z) = A(z) - \sum_{j=1}^{n} B_j(z) D_j(z)^{-1} C_j(z)$ where A, B_j, C_j, D_j, $j = 0, 1, ..., n$, are all polynomial.

In each case, eigenvalues and eigenvectors are defined the same way as in Definition 1.2. Both the ordinary and generalized eigenvalue problems are *linear eigenvalue problems*.

It is well known (see [FT01]) that $Q(\lambda)v = 0$ with Q as defined above is equivalent to

$$\left(\lambda \begin{bmatrix} A_2 & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} -A_1 & -A_0 \\ I & 0 \end{bmatrix}\right)\begin{bmatrix} \lambda v \\ v \end{bmatrix} = 0, \tag{1.3}$$

by which we turn the quadratic eigenvalue problem into a generalized eigenvalue problem using a companion linearization. The same process can be applied to any polynomial eigenvalue problem [MMMM06, GLR09], and more-

over different choices of polynomial basis lead to different kinds of linearization [ACL09] (e.g. the colleague linearization used in [EK12] and [BH13]). Since quadratic and polynomial eigenvalue problems can be rewritten as generalized eigenvalue problems, we call them *linearizable*. Rational eigenvalue problems are also *linearizable* [SB11].

There are also instances of nonlinear eigenvalue problems which we can rewrite as linearizable eigenvalue problems using a change of variable. For instance, $T(z) = A_2 z + A_1 \sqrt{z} + A_0$ can be transformed into a quadratic matrix-valued function $S(w) = A_2 w^2 + A_1 w + A_0$ with the change of variable $z = w^2$, and every eigenvalue of T will be the square of some eigenvalue of S.[4] Therefore we call the problem of finding eigenvalues of such a T a *linearizable eigenvalue problem* as well.[5] In contrast, we have the following definition.

Definition 1.4 (Genuinely nonlinear)**.** *A matrix-valued function $T(z)$ is called genuinely nonlinear if the problem of finding its eigenvalues is not linearizable. The eigenvalue problem for T is then called non-linearizable.*

Throughout this thesis we use classic notions from matrix theory and their extensions in terms of matrix-valued functions. One of these is the notion of pseudospectra. The classic definition of the ε-pseudospectrum of a matrix A (see [TE05]), denoted by $\sigma_\varepsilon(A)$, is the union $\bigcup_{\|E\|<\varepsilon} \sigma(A+E)$, where $\sigma(M)$ denotes the spectrum of a matrix M. Equivalently, $\sigma_\varepsilon(A) = \{z \in \mathbb{C} : \|zI - A\|^{-1} > \varepsilon^{-1}\}$. Thus, the following definition for the ε-pseudospectrum of a matrix-valued function T reduces to the usual notion of pseudospectrum in case $T(z) = zI - A$. See

[4] The converse may not hold, depending on whether the eigenvalues of S are in the range of the appropriate branch of the square root function.

[5] This is where we differ from other authors in our use of the term. Other authors would include $T(z) = A_2 z + A_1 \sqrt{z} + A_0$ in the category of genuinely nonlinear problems, as in [Sch08, p. 2].

Section 2.3 for more details.

Definition 1.5 (Pseudospectra)**.** *The ε-pseudospectrum of a matrix-valued function T with domain Ω is*

$$\Lambda_\varepsilon(T) = \{z \in \Omega : \|T(z)^{-1}\| > \varepsilon^{-1}\} \tag{1.4}$$

where $\|T(\lambda)^{-1}\|$ is considered infinite if $\lambda \in \Lambda(T)$. Equivalently, $\Lambda_\varepsilon(T)$ is the union $\bigcup_{E\in\mathcal{E}} \Lambda(T+E)$ where $\mathcal{E}$ is the set of matrix-valued functions on Ω with norm less than ε. Yet another equivalent definition is the union $\bigcup_{E\in\mathcal{E}_0} \Lambda(T+E)$ where $\mathcal{E}_0$ is the set of constant $n \times n$ matrices with norm less than ε.

The last set of terms we must set out have to do with *localization*, the process of deducing sets where the eigenvalues of a particular matrix-valued function T can be (or cannot be) without actually computing the eigenvalues. For example, the eigenvalues of a matrix A can be *localized* using Gershgorin's theorem from linear algebra, because Gershgorin's theorem shows that $\sigma(A)$ must lie in a certain union of disks in the complex plane whose centers and radii are computed from the elements of A.

If $U \subset \mathbb{C}$ is a subset that has been determined to contain no eigenvalues of a matrix-valued function T, then U is an *exclusion region for the spectrum of T*. Similarly, if U is a subset of the domain Ω of T that has been determined to contain all eigenvalues of T, then U is called an *inclusion region for the spectrum of T*. Sometimes we are only able to deduce inclusion regions within a proper subset of Ω. If $V \subset W \subset \Omega$ and it has been determined that any eigenvalues in W must lie in its subset V, then V is called an *inclusion region for the part of the spectrum of T within W*. Inclusion and exclusion theorems rely on conditions called *nonsingularity tests*, which determine regions where $T(z)^{-1}$ exists.

We will call theorems about localization *localization theorems*, *localization results*, or sometimes *inclusion theorems/results*. A theorem that splits Ω into a disjoint union of inclusion and exclusion regions will be referred to as a *global-type localization result*. Theorems that are guaranteed to do the same only for certain specified subsets of Ω will be called *regional-type localization results*. Finally, a result that pertains only to one eigenvalue or one cluster of eigenvalues at a time will be called a *local result*. All of our localization theorems also provide a way to count how many eigenvalues are in each component of the inclusion region under certain conditions. We will call this aspect of a localization theorem the *counting result*, as in "the counting result in Theorem 2.1."

1.3 A brief history of localization

Perhaps the first localization results were for analytic functions and polynomials. The most famous example is Rouché's theorem [Rou62, Théorème III], published in 1862. In its modern form, Rouché's theorem states that two functions f and $g := f + h$, holomorphic on and inside a simple closed curve $\Gamma \subset \mathbb{C}$, have the same number of zeros inside Γ if on Γ both $|h| < |f|$ and f is nonzero. For instance, using Rouché's theorem one can show that $z = 2 - e^{-z}$ has exactly one root with positive real part by taking $f(z) = z - 2$, $h(z) = e^{-z}$, and $\Gamma = i[-R, R] \cup \{z : |z| = R, \Re z > 0\}$ for arbitrarily large $R > 2$ [SS03, Exercise 6.7.10]. This means that all but one of the roots is localized to the left half-plane. Similarly, results due to Cauchy in 1829 and Pellet in 1881 [Mel13] allow one to localize the zeros of polynomials to certain disks and exclude them from certain annuli. And Mosier [Mos86] in 1986 defined so-called root neighborhoods of a polynomial p, easily computed regions with non-circular shapes, that allow one

to obtain approximations and counts for the roots of polynomials q which are sufficiently close to p in the proper sense.[6] Polynomial root localization theorems such as these are used in the program `MPSolve` [BNS13, BF00] released in 2000 by Bini and Fiorento, which performs simultaneous computation of polynomial roots.

A matrix approach to computing the roots of a polynomial, say, $p(z) = a_0 + a_1 z + a_2 z^2 + z^3$ is to write $p(\lambda) = 0$ in the form

$$\underbrace{\begin{bmatrix} -a_2 & -a_1 & a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\text{companion matrix}} \begin{bmatrix} \lambda^2 \\ \lambda \\ 1 \end{bmatrix} = \lambda \begin{bmatrix} \lambda^2 \\ \lambda \\ 1 \end{bmatrix}. \tag{1.5}$$

Thus, the roots of p are the eigenvalues of the companion matrix.[7] This is the formulation used for computing polynomial roots in the program `Eigensolve` [For, For02] and in MATLAB's `roots`. `Eigensolve` depends heavily on easy-to-compute eigenvalue localization regions obtained using Gershgorin's Theorem [Var04]. Using Gershgorin's theorem to localize polynomial roots is also done in [Car91], [Els73] among other work.

Gershgorin published his famous and elegant localization result in 1931 [Ger31].[8] Several others before him had come shockingly close to proving the same thing, but stopped short. For example, the underpinning of the Gershgorin theorem is a nonsingularity test based on diagonal dominance ([Ger31,

[6] Root neighborhoods are the polynomial analog of pseudospectra for matrices.

[7] Technically it's possible to reverse this process and compute or localize eigenvalues of a matrix A by localizing the roots of its characteristic polynomial $\det(zI - A)$. This is not done in practice because of numerical difficulties. Finding the coefficients in terms of sums and differences of traces or determinants [Pen87] is prone to cancellation [GL96, §2.4.3], and other methods [RI11b] have their own problems as well. In fact, the characteristic polynomial is usually computed by first computing the eigenvalues of A ([RI11a], also MATLAB's `poly`), which defeats the purpose of the exercise.

[8] We will discuss Gershgorin's theorem in more detail in Section 2.1.1.

Satz I], [Var04, Theorem 1.4]), and the 1881 article [Lév81] of Lucien Lévy, which Gershgorin cited, had already used a version of diagonal dominance. However, Lévy only used the concept to prove that a specific matrix arising in electrostatics can always be inverted. Minkowski used a version of diagonal dominance as well, for real matrices with all negative entries except for a positive diagonal, as the lemma of his 1900 article on algebraic number theory [Min00, Hülfssatz]. He, too, appears to have only concerned himself with showing that a particular matrix has nonzero determinant.

Others who came close to Gershgorin's theorem approached it from a different side. The usual proof, and the one Gershgorin himself gave, is to write out $(A - \lambda I)x = 0$ as a system of equations, move the diagonal term in each equation to the other side, take absolute values, and reason about the sizes of the terms. In contrast to this simplicity, previous notable localization results for matrix eigenvalues were encumbered by restrictive hypotheses. For instance, Bendixson proved in 1900 [Ben02] that if A is a real $n \times n$ matrix, then all of its eigenvalues live in the rectangle $[m, M] \times [-G, G] \subset \mathbb{C}$, where $G = \frac{1}{2}\sqrt{\frac{n(n-1)}{2}} \max_{j,k} |A_{jk} - A_{kj}|$ and m and M are the smallest and largest eigenvalues of $(A + A^T)/2$ (which are all real).[9] And in 1902, part of a letter from Hirsch to Bendixson was printed [Hir02], in which was contained an extension to matrices with complex entries. These results of Bendixson and Hirsch, which were proved by writing out the system of equations $(A - \lambda I)x = 0$ (with subsequent appeals to algebraic tricks and the theory of quadratic forms), clearly foreshadowed the Gershgorin theorem. Unfortunately, they were more complicated to state and to prove, and not even as useful.[10]

[9] This article also contains analogous localization results for eigenvalues of the pencil (A, B) for $A \in \mathbb{R}^{n \times n}$ and B symmetric positive definite.

[10] See the references in Householder [Hou64, §3.4] for an overview of eigenvalue localization pre-1964. Also see Varga's 2004 book [Var04] and the important 1960 article by Bauer and

The problem with modern-day localization results for matrix-valued functions has been similar; most of them (until [BH13], on which Chapter 2 of this thesis is based) apply only to a limited class of matrix-valued functions or are complicated to state and prove.[11] For instance, by using generalizations of Pellet's theorem mentioned above, eigenvalues for matrix polynomials were restricted to disks and excluded from annuli by Bini, Noferini and Sharify [BNS13] and by Melman [Mel13] in 2013. These rely on computing roots of univariate polynomials related to the given matrix polynomial, and thus do not extend to non-polynomial matrix-valued functions. Also, in 2003, Chu published Bauer-Fike-inspired results bounding the spectral variation (a distance measure between two sets of eigenvalues) between a given matrix polynomial and a perturbation [Chu03].[12] The simplest of these results bounds the distance between eigenvalues of $L(z) = Iz^\ell + \sum_{j=0}^{\ell-1} A_j z^j$ and those of its perturbation $\tilde{L}(z) = Iz^\ell + \sum_{j=0}^{\ell-1}(A_j + \delta A_j)z^j$ in terms of a Jordan triple for L and the perturbation size, measured as $\| [\delta A_0, ..., \delta A_{\ell-1}] \|$ [Chu03, Theorem 4.1]. However, objects like the Jordan triple are not available for matrix-valued functions in general, so these results do not generalize well either.

Although it may be expensive, pseudospectra can be used for localization as well. Recall that for a square matrix A and a chosen induced norm $\| \cdot \|$, the ε-pseudospectrum of A is the union $\bigcup_{\|E\|<\varepsilon} \sigma(A + E)$. Clearly then the spectrum of A is contained in $\sigma_\varepsilon(A)$ for every ε. The interesting part is that each connected component of $\sigma_\varepsilon(A)$ must contain at least one eigenvalue of A [TE05,

Fike [BF60] for norm-based inclusion regions. For inclusion regions based on pseudospectra, see [TE05, Theorem 2.4(i)].

[11] This is not to say that any localization theorem is useless; to the contrary, since the intersection of localization regions is a localization region, the more types we can use, the better.

[12] Interestingly, the author states that when he published these results in a technical report in 1992, he considered them of "negligible interest, due to the lack of applications." That has since changed.

Theorem 2.4(i)]. Since the study of pseudospectra came to the fore after the advent of computers, the field has only been around since the mid 1970s, with most work occurring in the 1990s and afteward.[13] It seems that the result about each component containing at least one eigenvalue first appeared in the 2001 article [ET01] by Embree and Trefethen.

Now, a matrix arising in applications may be constructed through a discretization process (say, the discretization of a differential operator, or the stiffness matrix in a finite element formulation), and thus its entries will have been computed numerically. From this perspective, the ε-pseudospectrum is not viewed so much as a localization tool, but as a means to understand the sensitivity of eigenvalues to ε-sized error in the computed entries. This perspective has inspired a whole class of generalizations of pseudospectra, and has been especially attractive because it allows structured perturbations. To see what is meant by this, notice that any matrix-valued function T can be written as a sum

$$T(z) = \sum_{j=1}^{m} p_j(z) A_j, \tag{1.6}$$

where each A_j is a constant matrix and each p_j is a function with the same domain as T. Thus, if T comes from an application where each A_j should be symmetric, then within the context of the application we need only worry about sensitivity of eigenvalues under errors in A_j that preserve its symmetry. In any case, by this thinking the functions p_j should remain untouched under perturbations. Thus, by 2001 a notion of pseudospectra tailored to the polynomial case $P(z) = \sum_{j=0}^{m} A_j z^j$, due to Tisseur and Higham, had appeared in [TH01].[14]

[13] See the bibliography at `http://www.cs.ox.ac.uk/pseudospectra/biblio.html`.

[14] There are also definitions that apply both to square and rectangular matrix polynomials, such as in [HT02]. These results give localization regions also, but it is not clear how to obtain eigenvalue counts for a given connected component of a localization region defined in this way.

The definition given was

$$\Lambda_\varepsilon(P) = \left\{\lambda \in \mathbb{C} : \left(\sum_{j=0}^{m}(A_j + \Delta A_j)\lambda^j\right)x = 0 \text{ for some } x \neq 0 \right. \\ \left. \text{and } \|\Delta A_j\| \leq \varepsilon\alpha_j,\ j = 0, 1, ..., m\right\}, \tag{1.7}$$

where the α_j are a chosen set of weights, or equivalently

$$\Lambda_\varepsilon(P) = \left\{z \in \mathbb{C} : \|P(z)^{-1}\| \geq (\varepsilon p(|\lambda|))^{-1}\right\}, \quad p(z) = \sum_{j=0}^{m}\alpha_j z^j. \tag{1.8}$$

More recently, a definition suitable in the context of time-delay problems was introduced in the 2006 article [GW06] by Green and Wagenknecht, having the form

$$\Lambda_\varepsilon(T) = \left\{z \in C : \det(zI - (A_0 + B_0) - \sum_{i=1}^{m}(A_i + B_i)\exp(-\tau_i z)) = 0 \right. \\ \left. \text{for some set of } B_i \text{ all satisfying } \|B_i\| \leq \varepsilon w_i,\ i = 0, 1, ..., m\right\}, \tag{1.9}$$

equivalent to

$$\Lambda_\varepsilon(T) = \left\{z \in \mathbb{C} : \|T(z)^{-1}\| \geq (\varepsilon g(z))^{-1}\right\}, \quad g(z) = w_0 + \sum_{j=1}^{m} w_j|\exp(-\tau_j z)|, \tag{1.10}$$

for matrix-valued functions $T(z) = zI - A_0 - \sum_{j=1}^{m} A_j e^{-\tau_j z}$. In both of these tailored definitions, the ε-pseudospectrum clearly contains the eigenvalues of their respective problems, and hence they are localization regions as well (although not used as such since there is no perceivable advantage).

Note that $zI - A_0 - \sum_{j=1}^{m} A_j e^{-\tau_j z}$ and matrix polynomials $A_0 + A_1 z + ... + A_m z^m$ are both special cases of the expansion $T(z) = \sum_{j=1}^{m} A_j p_j(z)$ written earlier. So the definition presented in [MGWN06, §2] by Michiels, Green, Wagenknecht and Niculescu in 2006 (also in [MN07b, Theorem 2.2]), of the form

$$\Lambda_\varepsilon(T) = \left\{z \in \mathbb{C} : \|T(z)^{-1}\|_\alpha \cdot \left\|\left[\frac{p_0(z)}{w_0} \dots \frac{p_m(z)}{w_m}\right]^T\right\|_\beta > \varepsilon^{-1}\right\}, \quad T(z) = \sum_{i=0}^{m} A_i p_i(z) \tag{1.11}$$

should not come as a surprise. This definition of ε-pseudospectrum for a general matrix-valued function T gives inclusion regions for the eigenvalues of T, but again it is hard to see how to use it effectively.

The main drawback to using any of the aforementioned generalizations of pseudospectra in localization is that (generically) no ε-pseudospectrum thus defined is guaranteed to contain eigenvalues of a problem that is easier to solve. In contrast, the definition of pseudospectra given in Definition 1.5 allows arbitrary changes to the matrix-valued function T under consideration. In terms of the expansion given above, this means we can perturb the p_j as well as the A_j, and make use of polynomial or rational approximants across arbitrarily large regions. Consequently, the pseudospectral localization result we present in Chapter 2 (also [BH13]) makes it easy to find eigenvalue approximations and counts, which would be difficult if not impossible to do with the definitions of pseudospectra that only allow perturbations to the coefficient matrices.

As far as we are aware, there are only two generalizations of pseudospectra in previous work that have the potential to be effectively used for localization, although no one seems to have pursued this direction. The first is in [WMG08, Def. 1], published in 2008 by Wagenknecht, Michiels, and Green, where perturbations to individual entries of $T(z)$ are allowed. However, the definition is more laborious to state than ours, and its peculiarities seem to add no theoretical value as far as localization goes. The second is the relatively old generalization

$$\Lambda_\varepsilon(T) = \left\{ z \in \mathbb{C} : \|T(z)\|_2 \, \|T(z)^{-1}\|_2 \geq \varepsilon^{-1} \right\} \tag{1.12}$$

published in [CR01] by Cullum and Ruehli in 2001, apparently before the trend was to focus on structured pseudospectra. This definition can be used for a localization result as well, but the factor of $\|T(z)^{-1}\|_2$ is an encumberance rather

than a help; in our framework, it obstructs the process of comparing T to a simplified function and hence makes eigenvalue counts more complicated to obtain.

As far as non-pseudospectral, global-type localization results that work for general matrix-valued functions, besides the results presented in [BH13] and Chapter 2 of this thesis we know of only two, and even those are rather restrictive. First, Jarlebring showed in his thesis in 2008 that if a matrix-valued function has a certain contractive property then a Bauer-Fike-related localization result could be derived using a fixed point iteration [Jar08, Theorem 4.20]. For example, he used his theorem to show the eigenvalues of $G_2(s) = A_1 + A_2\cos(|s|)$ (in the sense that $G_2(s) - sI$ is singular) are contained in disks of radius 0.223 centered at the eigenvalues of A_1, where

$$A_1 = \begin{bmatrix} 7 & 10 \\ 8 & -15 \end{bmatrix}, \qquad A_2 = \frac{1}{10}\begin{bmatrix} -1 & -1 \\ 0.7 & 0.6 \end{bmatrix}. \tag{1.13}$$

The second was developed for matrix-valued functions such as $z \mapsto zI - A_0 - A_1e^{-z}$ arising in the study of delay differential equations. It is possible to find an "envelope curve" $|z| = \|A_0\|_2 + \|A_1\|_2 e^{-\Re\mathrm{e}\, z}$ which is to the right of all eigenvalues of T (see the 2007 book [MN07b, Prop. 1.10] by Michiels and Niculescu and a similar result in [MN07b, §1.2] for neutral type equations). This is generalizable to general matrix-valued functions $T(z)$ by writing $T(z) = zI - (T(z) - zI)$ and deducing that any eigenvalue λ of T must satisfy $|\lambda| \leq \|T(\lambda) - \lambda I\|$. Splitting the right-hand side appropriately gives something similar to the envelope curve. However, this result does not provide eigenvalue approximations or counts.

Before concluding this section it is worth mentioning some local analysis that has been done for eigenvalues of general matrix-valued functions (including the

class of genuinely nonlinear ones). First of all, a matrix version of Rouché's theorem [GGK90, Theorem 9.2, p. 206] is used in [MGWN06, Prop. 2] to prove that if T is a matrix-valued function and $\mathcal{D} \subset \mathbb{C}$ is a given disk, then every perturbation of T, provided it is small enough, has the same number of eigenvalues in $\mathcal{D}$ as T does. This is used to prove continuity of the eigenvalues with respect to perturbations, but could also be used to obtain eigenvalue counts within a particular disk $\mathcal{D}$ of interest through comparison to a simplified problem. Second, although these are by no means concrete localization results, there has been much attention paid to perturbation theory for eigenvalues, especially for polynomial eigenvalue problems. Work since 2000 on backward error analysis and eigenvalue condition numbers includes [Tis00], [DT03], [HLT07], [Bor10], and [AM11]. The pseudospectra definitions addressed above also assist in this type of perturbation analysis.

1.4 Contributions and outline

The first contribution of this thesis is theorems for localizing eigenvalues of general matrix-valued functions $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$. The most important one is Theorem 2.1[15], a generalization of Gershgorin's theorem [Var04, Theorem 1.1]. Just like Gershgorin's theorem, Theorem 2.1 is both very easy to use and very useful; it is a localization theorem of global type, allowing Ω to be split into a disjoint union of inclusion and exclusion regions, and visualizing these regions requires only basic arithmetic and a contour plotter. The next most important theorem is a localization result based on a nonlinear generalization of pseudospectra; The-

[15] A slight generalization of this theorem also appears as [BH13, Theorem 3.1].

orem 2.5[16] allows us to obtain concrete inclusion regions and eigenvalue approximations by comparison with a problem that is easier to solve. A close third is a block version of our Gershgorin generalization, Theorem 2.2. Although clumsier to use than Theorem 2.1, there are times it works when Theorem 2.1 is of no help. We also include some Bauer-Fike-inspired theorems for matrix-valued functions of the form $T(z) = A - zI + E(z)$, most useful when $E : \Omega \to \mathbb{C}^{n\times n}$ is small in norm over subregions of Ω.

What makes Theorems 2.1, 2.2, and 2.5 so special is how generally they can be applied and how little work it takes to do so. We hope this will be made abundantly clear by the gallery of examples presented in Chapter 3. Through experience, we have found that there are several basic strategies for getting the most out of these theorems, and have grouped the problems in the chapter by strategy. In some cases we do not stop at localization, but proceed to compute eigenvalues using the initial localization step to make the computations easier and more reliable. In particular, we

- choose good initial guesses for Kressner's block Newton algorithm [Kre09],
- choose good contours for Beyn's integral-based method [Bey12], and
- validate the computed eigenvalues using the counting results in the theorems.

In cases where we are able to localize eigenvalues to tiny regions, we also use Theorem 2.1 as a theoretical tool to analytically derive concrete error bounds between computed eigenvalue approximations and the true values.

[16] Appears also as [BH13, Theorem 4.4].

The next contribution is to apply our notion of pseudospectra to matrix-valued functions arising in higher order and delay differential equations. Rather than viewing pseudospectra as merely relating to the size of the perturbation that will knock a system out of stability, we follow [TE05] and use our definition to study *transient growth*, i.e., solution behavior for intermediate times, before it ultimately settles down into the asymptotic behavior dictated by eigenvalues. We are able to concretely bound transient behavior both above and below, and a model for a semiconductor laser with phase-conjugate feedback acts as the central example.

The last contribution of this thesis is about the resonances for electrons trapped in circular quantum corrals, microscopic structures built by placing atoms in a circle on a metal surface. Starting from a differential equation, we reduce the problem of finding resonances to a nonlinear eigenvalue problem for a matrix-valued function, discussing the error in each step. Along the way, we use the pseudospectral inclusion result Theorem 2.5 and the Gershgorin generalization Theorem 2.1 in both theoretical and computational capacities. Theorem 2.5 and first order perturbation theory results are then used to create a framework for *comparing* models. We then use this framework to bound the error between resonances computed from the naïve particle-in-a-box model and resonances computed from a model that takes quantum tunneling into account.

CHAPTER 2

LOCALIZATION THEOREMS FOR GENERAL MATRIX-VALUED FUNCTIONS[1]

2.1 Gershgorin theorems

In this section, we show how to localize eigenvalues of a matrix-valued function T by looking at simple functions of its entries T_{ij}.

2.1.1 Inspiration

Suppose $A \in \mathbb{C}^{n\times n}$ is split into $A = D + F$ where D is diagonal. Since the zeros of a polynomial depend continuously on the polynomial coefficients, the zeros of $\det(D + sF - zI)$ move continuously from the diagonal entries of D to the eigenvalues of A as s is increased from 0 to 1.

We can restrict where the paths between the eigenvalues of D and A can be in the following way. First, let s be arbitrary and let λ, v be an eigenpair for $M = D + sF$, that is, λ is on one of the paths connecting eigenvalues of D and A. Then $\sum_{j=1}^{n} M_{ij}v_j = \lambda v_i$ for all $i = 1, ...n$. A little manipulation turns this into $\sum_{j=1}^{n} sF_{ij}v_j = (\lambda - D_{ii})v_i$, $i = 1, ..., n$. Then $|\lambda - D_{ii}|\,|v_i| \leq s\sum_{j=1}^{n} |F_{ij}|\,|v_j|$ for all $i = 1, ..., n$. Hence, $|\lambda - D_{ii}| \leq s\sum_{j=1}^{n} |F_{ij}|$ for at least one value of i, namely the one such that $|v_i| = \max_j |v_j|$. Furthermore, for this i, $|\lambda - D_{ii}| \leq \sum_{j=1}^{n} |F_{ij}|$ since $0 \leq s \leq 1$.

The preceding are the proof ingredients for the Gershgorin Circle Theorem [Var04], which states that the eigenvalues of A live in the union of the n

[1] This chapter is partially based on [BH13].

Gershgorin disks $\{|z - D_{ii}| \leq \sum_j |F_{ij}|\}$, $i = 1, 2, ..., n$, and that in every connected component of this union the number of eigenvalues of D and A are the same.

Remark 2.1. The eigenvalues of a matrix A can be localized by applying the Gershgorin Circle Theorem to any matrix with the same eigenvalues, namely VAV^{-1} for nonsingular V. As discussed in [Var04, §1.4], this could be useful if the Jordan normal form of A is expensive or inaccurate to compute.

Remark 2.2. Also as discussed in [Var04, §1.1], the Gershgorin Circle Theorem is just one example of how to turn a nonsingularity test into an eigenvalue localization result.

A major convenience of the Gershgorin Circle Theorem is that each Gershgorin disk is computed merely by adding up a few scalars. The next best thing is localization regions that come from analyzing small matrices. For instance, if we take a square matrix A with $Av = \lambda v$ and partition this equation into m block rows and columns we get

$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ A_{21} & A_{22} & \dots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mm} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} = \begin{bmatrix} \lambda v_1 \\ \lambda v_2 \\ \vdots \\ \lambda v_m \end{bmatrix}. \tag{2.1}$$

In parallel to the argument used for the Gershgorin Circle Theorem, let us split A into $D + F$, where D is block diagonal according to the above partition. Then the i-th block row equation is $\sum_{j=1}^{m} F_{ij}v_j + D_{ii}v_i = \lambda v_i$, which we can rearrange as $(D_{ii} - \lambda I)v_i = -\sum_{j=1}^{m} F_{ij}v_j$. Then for $\|v_i\| = \max_j \|v_j\|$,

$$\frac{\|(D_{ii} - \lambda I)v_i\|}{\|v_i\|} \leq \sum_{j=1}^{m} \|F_{ij}\|. \tag{2.2}$$

By definition of induced norm, the left hand side is greater than or equal to $\|(D_{ii} - \lambda I)^{-1}\|^{-1}$. Therefore, according to the block diagonal version of Gershgorin's theorem [Var04, Theorem 6.3], eigenvalues of A live in the union of the m Gershgorin regions $\{z \in \mathbb{C} : \|(D_{ii} - zI)^{-1}\|^{-1} \leq \sum_{j=1}^{m} \|F_{ij}\|\}$, $i = 1, 2, ..., m$.

Remark 2.3. As z approaches an eigenvalue of D_{ii}, $\|(D_{ii} - zI)^{-1}\|^{-1}$ approaches zero, so we should consider the eigenvalues of each D_{ii} to live in the i-th set. Furthermore, each connected component of the union should contain the same number of eigenvalues of D and A by the same reasoning as before.

2.1.2 Useful tools

Most of the proof ingredients in Section 2.1.1 will essentially be reused to derive localization results for the eigenvalues of general matrix-valued functions $T(z)$. However, there are two points that must be handled with more care in the latter case. First, in the previous discussion of the Gershgorin Circle Theorem, we used the fact that the zeros of a polynomial are continuous functions of the coefficients to justify that the same number of eigenvalues of D and A live in a given connected component of the union of Gershgorin regions. Since $\det T(z)$ is not necessarily a polynomial, we will need the argument principle to justify this counting result. Second, in contrast to Gershgorin disks for a matrix A, the "Gershgorin regions" we will derive for the eigenvalues of $T(z)$ can be arbitrarily shaped and we would like to be able to say that every connected component contains at least one eigenvalue of T. While this is completely obvious for Gershgorin disks, the theory of subharmonic functions will be needed to show it for the Gershgorin regions for T.

Recall that if f is a function analytic in open set U, then for Γ a simple closed curve inside U on which f is nonzero, the number of zeros (counting multiplicity) of f inside Γ is

$$N_f = \frac{1}{2\pi i} \oint_\Gamma \frac{f'(z)}{f(z)}\, dz. \tag{2.3}$$

This is usually called the argument principle and can be found in standard complex analysis references, e.g., [Rud87, Theorem 10.43a]. For a given regular, analytic matrix-valued function $T(z)$ with domain $\Omega \supset U$, the number of eigenvalues of T (counting multiplicity) inside Γ can be computed by putting $f(z) = \det T(z)$.[2]

Lemma 2.1. *(based on [BH13, Lemma 2.1]) Suppose $T : \Omega \to \mathbb{C}^{n\times n}$ and $E : \Omega \to \mathbb{C}^{n\times n}$ are analytic, and that $\Gamma \subset \Omega$ is a simple closed contour. If $T(z) + sE(z)$ is nonsingular for all $s \in [0, 1]$ and all $z \in \Gamma$, then T and $T + E$ have the same number of eigenvalues inside Γ, counting multiplicity.*

Proof. Define $f(z; s) = \det(T(z) + sE(z))$. By hypothesis, $f(z; s) \neq 0$ on $\Gamma \times [0, 1]$ and we can apply the argument principle to compute $N_{f(\cdot,s)}$ for every $s \in [0, 1]$. Since $N_{f(\cdot,s)}$ is integer-valued and continuous in s, it is constant. □

Recall that a continuous function $f : \Omega \subset \mathbb{C} \to \mathbb{R}$ is called subharmonic [Rud87, Definition 17.1] if it satisfies

$$f(a) \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} u(a + re^{i\theta})\, d\theta \tag{2.4}$$

whenever the integrand lies entirely in Ω. It is easy to see from this property that subharmonic functions obey the maximum principle, i.e., the maximum over a set cannot occur on the interior or (2.4) would be violated.

[2] Since $\frac{d}{dz} \det(T(z)) = \det(T(z)) \operatorname{tr}\left(T(z)^{-1}T'(z)\right)$ [MN07a, III.8.3], $N_{\det T(z)} = \frac{1}{2\pi i} \oint_\Gamma \operatorname{tr}\left(T(z)^{-1}T'(z)\right) dz$.

Lemma 2.2. *If $v : \Omega \to \mathbb{C}^n$ is an analytic vector-valued function, $v \neq 0$ on K, then $\|v\|_1$ is subharmonic, where $\|v\|_1 = \sum_{j=1}^n |v_j(z)|$.*

Proof. Since each $v_j(z)$ is analytic, each $\log|v_j(z)|$ is harmonic (this can be verified directly using the fact that v_j satisfies the Cauchy-Riemann equations) and therefore trivially subharmonic. Second, the composition $|v_j(z)| = \exp(\log|v_j(z)|)$ is subharmonic [Rud87, Theorem 17.2] because exp is monotonically increasing and convex (preserving (2.4) by Jensen's inequality [Rud87, Theorem 3.3]). Hence, each $|v_j(z)|$ must obey (2.4), i.e.,

$$|v_j(a)| \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |v_j(a + re^{i\theta})|\, d\theta. \tag{2.5}$$

Furthermore, $\|v(z)\|_1$ is continuous since each $|v_j(z)|$ is continuous. Finally,

$$\|v(a)\|_1 = \sum_{j=1}^n |v_j(a)| \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\sum_{j=1}^n |v_j(a + re^{i\theta})| \right) d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \|v(a + re^{i\theta})\|_1\, d\theta \tag{2.6}$$

shows that $\|v(z)\|_1$ is subharmonic, as desired. □

That was basically a warm-up, but it is all we need for the generalization of the Gershgorin Circle Theorem to the case of matrix-valued functions. For the block diagonal extension, we need a more general version of the previous lemma, and a lemma that involves matrix norms.

Lemma 2.3. *If $v : K \to \mathbb{C}^n$ is an analytic vector-valued function on a compact set K, with $v \neq 0$ on K, then $\varphi(z) = \|v(z)\|$ is subharmonic for any norm $\|\cdot\|$.*

Proof. First, notice that $\varphi(z)$ is continuous because v is. Next, as a consequence [Rud87, Remarks 5.21] of the Hahn-Banach theorem, $\|v(z)\| = \sup_{\ell^* \in \mathcal{B}^*} |\ell^* v(z)|$ where $\mathcal{B}^*$ is the set of bounded linear functionals on $\mathbb{C}^n$ with norm

equal to 1. For any given $\ell^* \in \mathcal{B}^*$, $\ell^* v(z)$ is an analytic function $\mathbb{C} \to \mathbb{C}$. Therefore $|\ell^* v(z)|$ is subharmonic by the same reasoning as in Lemma 2.2. We will now find a monotone increasing sequence of continuous, subharmonic functions that converges uniformly to φ on K.

Since the dual $\mathcal{B}^*$ is finite-dimensional, it has a countable dense subset (ℓ_j). Define $\varphi_m(z) = \max_{j \le m} |\ell_j^* v(z)|$. Because f, g continuous and subharmonic implies $\max(f, g)$ is continuous ($\max(f, g) = \frac{1}{2}(f + g + |f - g|)$) and subharmonic (check (2.4)), it follows that each φ_m is continuous and subharmonic. Since $\varphi_m(z) \le \varphi_{m+1}(z)$ for all m and all $z \in K$ and $\varphi_m \to \varphi$ pointwise, $\varphi_m \to \varphi$ uniformly on K [Rud76, Theorem 7.13]. Using the fact that each φ_m satisfies (2.4), we have

$$\varphi(a) = \lim_{m \to \infty} \varphi_m(a) \le \lim_{m \to \infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_m(a + re^{i\theta})\, d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi(a + re^{i\theta})\, d\theta \tag{2.7}$$

where the last equality follows from the uniform convergence $\varphi_m \to \varphi$ by [Rud76, Theorem 7.16]. Therefore φ is subharmonic. □

Our last lemma of this type has to do with norms of matrix-valued functions.

Lemma 2.4. *Suppose $B : K \to \mathbb{C}^{p \times n}$ is a regular, analytic matrix-valued function on compact domain K, $B \neq 0$ on K. Then $\varphi(z) = \|B(z)\|$ is subharmonic for any norm $\|\cdot\|$ induced by a vector norm.*

Proof. Since $\varphi(z) = \max_{\|v\|=1} \|B(z)v\|$, let (v_j) be a countable dense subset of the unit ball of $\mathbb{C}^n$ under $\|\cdot\|$ such that $B(z)v_j \neq 0$ for all v_j. Define $\varphi_m(z) = \max_{j \le m} |B(z)v_j|$. Each $B(z)v_j$ is an analytic, nonzero, vector-valued function on K, hence each $\|B(z)v_j\|$ is continuous and subharmonic by Lemma 2.3. (φ_m) is pointwise monotone increasing and $\varphi_m \to \varphi$ pointwise, so $\varphi_m \to \varphi$ uniformly by [Rud76, Theorem 7.13], and hence $\|B(z)\|$ is subharmonic by [Rud76, Theorem 7.16]. □

2.1.3 Extensions for matrix-valued functions

In this section, for a given set S we denote by S_ε the union of closed ε-balls centered at elements of S, i.e., $S_\varepsilon = \bigcup_{z\in S}\{w \in \mathbb{C} : |w - z| \le \varepsilon\}$. The following extension to the Gershgorin Circle Theorem is adapted from [BH13, Theorem 3.1].

Theorem 2.1. *Suppose $T(z) = D(z) + E(z)$ where $D, E : \Omega \to \mathbb{C}^{n\times n}$ are analytic and D is diagonal. Define $G = \bigcup_{j=1}^n G_j$, where*

$$G_j = \Big\{z \in \Omega : |D_{jj}(z)| \le r_j(z)\Big\}, \qquad r_j(z) = \sum_{k=1}^{n} |E_{jk}(z)|. \tag{2.8}$$

is called the j-th Gershgorin region. Then

(a) $\Lambda(T) \subset G$.

(b) *If $\mathcal{U}$ is a bounded connected component of G such that the closure of $\mathcal{U}$ in $\mathbb{C}$ is contained in Ω, then $\mathcal{U}$ contains the same number of eigenvalues of T and D.*

(c) *If in addition $\mathcal{U}$ is the union $\bigcup_{k=1}^m \mathcal{U}_k$ where each $\mathcal{U}_k$ is a connected component of some Gershgorin region G_j, then $\mathcal{U}$ contains at least m eigenvalues.*

Proof. Following the proof sketch for the Gershgorin Circle Theorem in Section 2.1.1, suppose λ, v is an eigenpair for $D(z) + sE(z)$ for some $s \in [0, 1]$, i.e., $(D(\lambda) + sE(\lambda))v = 0$. Then for each row j, $\sum_{k=1}^n sE_{jk}(\lambda)v_k + D_{jj}(\lambda)v_j = 0$, or $D_{jj}(\lambda)v_j = -s\sum_{k=1}^n E_{jk}(\lambda)v_k$. For $|v_j| = \max_k |v_k|$, $|D_{jj}(\lambda)| \le r_j(\lambda)$. Therefore every eigenvalue of $D(z) + sE(z)$ is contained in some G_j, for every $s \in [0, 1]$. This statement with $s = 1$ is (a).

To prove (b), first we show that the distance between $\mathcal{U}$ and $G - \mathcal{U}$ is finite. Now, G is closed in Ω because it is the finite union of closed sets G_j. By definition

of connected component, G is the union of disjoint sets $\mathcal{U}$ and $G-\mathcal{U}$, from which it is clear that each is closed in Ω as well. Therefore $\mathcal{U}$ and $G-\mathcal{U}$ cannot share any limit points in Ω. By assumption, the closure of $\mathcal{U}$ in $\mathbb{C}$ is a subset of Ω, whereas the closure of $G-\mathcal{U}$ in $\mathbb{C}$ can only add limit points from $\mathbb{C}-\Omega$. Hence there is a finite distance between $\mathcal{U}$ and $G-\mathcal{U}$ in $\mathbb{C}$. Then it is clear that $\partial\mathcal{U}_\varepsilon \subset (\mathbb{C}-G)\cap\Omega$ for sufficiently small ε. Since $D(z)+sE(z)$ is nonsingular in $(\mathbb{C}-G)\cap\Omega$ for every s, we can apply Lemma 2.1 with $\Gamma = \partial\mathcal{U}_\varepsilon$ to show that D and T have the same number of eigenvalues inside each $\mathcal{U}_\varepsilon$ for every sufficiently small ε, and hence D and T have the same number of eigenvalues in $\mathcal{U}$.

To prove (c), it is enough to show that each bounded connected component $\mathcal{V}$ of G_j with $\mathcal{V}\subset\Omega$ contains at least one zero of the scalar function D_{jj}. First notice that $\mathcal{V}$ is closed by the same reasoning used previously. Therefore $\mathcal{V}$ is compact. Now define the meromorphic vector-valued function $v:\Omega\to\bar{\mathbb{C}}^n$ [3] by $v_k(z) = E_{jk}(z)/D_{jj}(z)$, so that $G_j = \{z\in\Omega : 1\le\|v(z)\|_1\}$. By previous argument, $\partial\mathcal{V}_\varepsilon\subset(\mathbb{C}-G_j)\cap\Omega$ for ε small enough. Notice that $\|v(z)\|_1<1$ on $\partial\mathcal{V}_\varepsilon$ but $\|v(z)\|_1\ge 1$ on $\mathcal{V}\subset\mathcal{V}_\varepsilon$. If $D_{jj}\neq 0$ on a given $\bar{\mathcal{V}}_\varepsilon$, which is compact, then Lemma 2.2 shows that $\|v(z)\|_1$ is subharmonic on $\mathcal{V}_\varepsilon$ and hence the maximum of $\|v(z)\|_1$ must occur on the boundary $\partial\mathcal{V}_\varepsilon$. But as we have just pointed out, this is impossible. Therefore D_{jj} must have at least one zero in each $\mathcal{V}_\varepsilon$ for every ε sufficiently small. Thus, D_{jj} must have at least one zero in $\mathcal{V}$. □

Remark 2.4. By applying permutations and similarity transformations that preserve the spectrum, different Gershgorin regions can be obtained. Taking the intersection of several is often helpful.

The proof of b) given above hinges on hypotheses that allow us to use

[3] The symbol $\bar{\mathbb{C}}$ denotes the extended complex plane $\mathbb{C}\cup\{\infty\}$.

Lemma 2.1. These hypotheses are in fact necessary, as demonstrated by the two simple examples that follow. The first shows what can happen if a component of the inclusion region G is not contained in Ω.

Example 2.1. Consider the matrix-valued function

$$T(z) = \begin{bmatrix} z - 0.2\sqrt{z} + 1 & -1 \\ 0.4\sqrt{z} & 1 \end{bmatrix}. \tag{2.9}$$

Since the square root function is multi-valued, the domain of T must exclude a branch cut in order that T be analytic on its domain Ω. Taking the principal branch of the square root corresponds to $\Omega = \mathbb{C} - (-\infty, 0]$. By Theorem 2.1, the eigenvalues of T are confined to $G_1 \cup G_2$, where

$$G_1 = \left\{ z : |z - 0.2\sqrt{z} + 1| \leq 1 \right\} \tag{2.10}$$

$$G_2 = \left\{ z : 1 \leq |0.4\sqrt{z}| \right\}. \tag{2.11}$$

These regions are shaded and labeled in Figure 2.1.

Now, $\det T(z) = z + 0.2\sqrt{z} + 1$, which is formally satisfied by $\sqrt{z} = -0.1 \pm i\sqrt{0.99}$. These numbers are in the left half-plane. However, the range of the principal branch of the square root is the right half-plane. Therefore T has no eigenvalues at all given our choice of square root branch. Taking $D(z) = \mathrm{diag}(z - 0.2\sqrt{z}+1, 1)$, the zeros of $\det D(z)$ satisfy $0.1 \pm i\sqrt{0.99}$, which *are* in the range of the principal branch of the square root. The corresponding values of z are $\approx -0.980 \pm 0.199i$, which are in the G_1 component. Hence D and T have different numbers of eigenvalues in the component G_1, but this does not violate Theorem 2.1(b) because G_1 intersects Ω^c.

The second example addresses the hypothesis of boundedness.

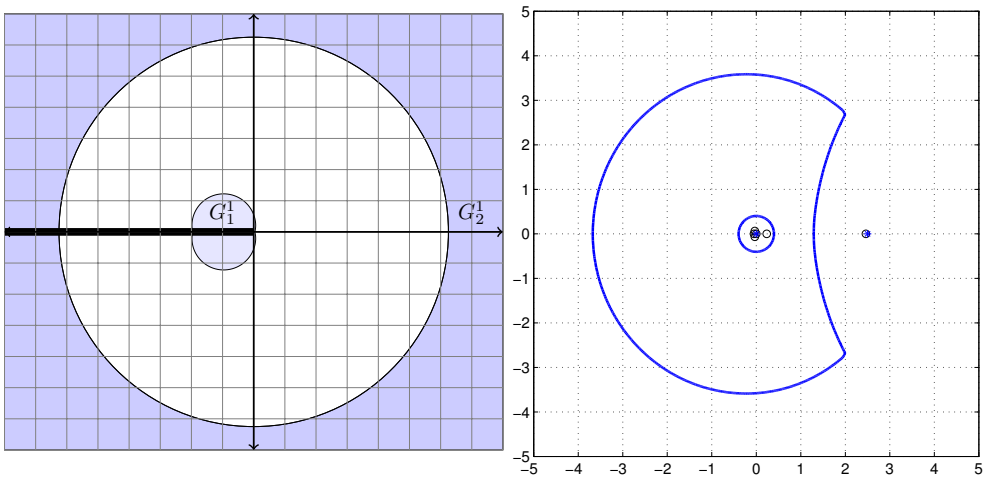

Figure 2.1: Inclusion regions for hypothesis-illustrating examples.

Example 2.2. The `bilby` example from [BHM+13] is a quadratic eigenvalue problem $T(z) = Az^2 + Bz + C$,

$$A = \begin{bmatrix} 0 & 0.05 & 0.055 & 0.08 & 0.1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.22 & 0 & 0 \\ 0 & 0 & 0 & 0.32 & 0.4 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0.01 & 0.02 & 0.01 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0.04 & -1 & 0 & 0 \\ 0 & 0 & 0.08 & -1 & 0 \\ 0 & 0 & 0 & 0.04 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 0.1 & 0.04 & 0.025 & 0.01 & 0 \\ 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0.16 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.04 & 0 \end{bmatrix}. \tag{2.12}$$

Eigenvalues of T are confined to $G = \bigcup_{j=1}^{5} G_j$,

$$G_1 = \left\{ z : |-z + 0.1| \le \sum_{k=2}^{5} |A_{1k}z^2 + B_{1k}z + C_{1k}| \right\} \tag{2.13}$$

$$G_2 = \left\{ z : |-z| \le 0.4 \right\} \tag{2.14}$$

$$G_3 = \left\{ z : |-z| \le |0.2z^2 + 0.04z + 0.16| \right\} \tag{2.15}$$

$$G_4 = \left\{ z : |-z| \le |0.22z^2 + 0.08z + 0.1| \right\} \tag{2.16}$$

$$G_5 = \left\{ z : |0.4z^2 - z| \le |0.32z^2 + 0.04z + 0.04| \right\}. \tag{2.17}$$

The boundary of the union G is depicted in Figure 2.1 (right) in thick blue line, along with eigenvalues of T ($\circ$) and the diagonal $D(z) = \mathrm{diag}([-z +$

$0.1, -z, -z, -z, 0.4z^2 - z]$) $(*)$. The Gershgorin region G itself is the union of the regions outside the outer curve and within the inner curve, thus the inner component must contain the same number of eigenvalues of T and D by Theorem 2.1(b). Computing the eigenvalues of T and D (e.g., with `polyeig` from MATLAB) confirms that each has 5 eigenvalues in the inner component. We also learn that D has only one eigenvalue in the outer component, at 2.5, while T has two: one at ≈ 2.5 and one at ≈ 1123. The fact that D and T have different numbers of eigenvalues in the outer component is not a violation of Theorem 2.1(b) because the outer component is unbounded.

The next theorem is an extension of a block version of Gershgorin's theorem [Var04, Theorem 6.3].

Theorem 2.2. *Suppose $T(z) = D(z) + E(z)$ where $D, E : \Omega \to \mathbb{C}^{n\times n}$ are analytic, D is block diagonal, and E is split into blocks according to the same partition of $\{1, 2, ..., n\}$ and the j,k block is size $n_j \times n_k$. Let $\|\cdot\|$ denote any vector norm and the corresponding induced matrix norm. Define*

$$\tilde{G}_j = \left\{ z \in \Omega : 1 \leq \sum_{k=1}^{n} \|D_{jj}(z)^{-1} E_{jk}(z)\| \right\}, \tag{2.18}$$

$$G_j = \left\{ z \in \Omega : \|D_{jj}(z)^{-1}\|^{-1} \leq \sum_{k=1}^{n} \|F_{jk}(z)\| \right\} \tag{2.19}$$

and $G = \bigcup_j G_j$, and $\tilde{G} = \bigcup_j \tilde{G}_j$. Then

(a) $\Lambda(T) \subset \tilde{G} \subset G$.

(b) *If $\mathcal{U}$ is a bounded connected component of G (or $\tilde{G}$) such that the closure of $\mathcal{U}$ in $\mathbb{C}$ is in Ω, then $\mathcal{U}$ contains the same number of eigenvalues of T and D.*

(c) *If in addition $\mathcal{U}$ is the union $\bigcup_{k=1}^{m} \mathcal{U}_k$ where each $\mathcal{U}_k$ is a connected component of some $\tilde{G}_j$, then $\mathcal{U}$ contains at least m eigenvalues.*

Proof. Suppose $(D(\lambda) + sE(\lambda))\, v = 0$ and split v accordingly so that v_k is length n_k. Then the j-th block row of this equation is $\sum_{k=1}^{n} sE_{jk}(\lambda)v_k + D_{jj}(\lambda)v_j = 0$, or equivalently $v_j = -s\sum_{k=1}^{n} D_{jj}(\lambda)^{-1}E_{jk}(\lambda)v_k$. If $\|v_j\| = \max_k \|v_k\|$, then $1 \leq \sum_{k=1}^{n} \|D_{jj}(\lambda)^{-1}E_{jk}(\lambda)\|$, which shows $\lambda \in \tilde{G}_j$. So for every $s \in [0,1]$, each eigenvalue of $D(z) + sE(z)$ is in some $\tilde{G}_j$. And since $1 \leq \sum_k \|D_{jj}(z)^{-1}E_{jk}(z)\|$ implies $1 \leq \sum_k \|D_{jj}(z)^{-1}\|\,\|E_{jk}(z)\|$ which is equivalent to $\|D_{jj}(z)^{-1}\|^{-1} \leq \sum_k \|E_{jk}(z)\|$, it follows that $\tilde{G}_j \subset G_j$ for each j. Therefore we have (a). The proof for (b) is the same as in Theorem 2.1.

To prove (c), it suffices to show that if $\mathcal{V}$ is a bounded connected component of $\tilde{G}_j$ with $\mathcal{V} \subset \Omega$, then $\mathcal{V}$ contains at least one eigenvalue of D_{jj}. As a closed and bounded set, $\mathcal{V}$ is compact. As in Theorem 2.1, $\partial\mathcal{V}_\varepsilon \subset (\mathbb{C} - \tilde{G}_j) \cap \Omega$ for ε small enough. Now, for each k, $D_{jj}(z)^{-1}F_{jk}(z)$ is a meromorphic matrix-valued function $\Omega \to \bar{C}^{n_j \times n_k}$ with poles at the eigenvalues of D_{jj}. If D_{jj} has no eigenvalues in $\bar{\mathcal{V}}_\varepsilon$, which is compact, then Lemma 2.4 shows that $\|D_{jj}(z)^{-1}F_{jk}(z)\|$ is subharmonic on $\mathcal{V}_\varepsilon$ for every $k = 1, 2, ..., n$, and hence $f(z) := \sum_{k=1}^{n} \|D_{jj}(z)^{-1}F_{jk}(z)\|$ is subharmonic on $\mathcal{V}_\varepsilon$. Then the maximum of f over $\mathcal{V}_\varepsilon$ must occur on the boundary. But this is impossible, because by definition of $\tilde{G}_j$ we have $f(z) \geq 1$ on $\mathcal{V} \subset \mathcal{V}_\varepsilon$ and $f(z) < 1$ on $\partial\mathcal{V}_\varepsilon$. So, the matrix-valued function D_{jj} must have at least one eigenvalue in $\mathcal{V}_\varepsilon$ for every sufficiently small ε. Therefore D_{jj} must have at least one eigenvalue in $\mathcal{V}$. □

2.2 Bauer-Fike theorems

Suppose $B = A + E \in \mathbb{C}^{n\times n}$ where $AV = VD$ is an eigendecomposition for A and E is small. If λ is an eigenvalue of B, then $(\lambda I - B)u = 0$ for some $u \neq 0$ by definition,

which can be rearranged as $(\lambda I - A)u = Eu$. Now, $(zI - D)^{-1} = \sum_{j=1}^{n}(z - D_{jj})^{-1}e_j e_j^T$ where e_j is the j-th column of the identity matrix. Similarly,

$$(zI - A)^{-1} = V(zI - D)^{-1}V^{-1} = \sum_{j=1}^{N}(z - D_{jj})^{-1}(Ve_j)(e_j^T V^{-1}). \tag{2.20}$$

Therefore $u = \sum_{j=1}^{N}(\lambda - D_{jj})^{-1}(Ve_j)(e_j^T V^{-1})Eu$. Taking norms, this implies that if λ is an eigenvalue of B, then

$$1 \leq \|E\| \sum_{j=1}^{N} |\lambda - D_{jj}|^{-1} \|(Ve_j)(e_j^T V^{-1})\|. \tag{2.21}$$

Similar to the derivation of Gershgorin regions, let

$$\|(Ve_i)(e_i^T V^{-1})\|/|\lambda - D_{ii}| = \max_j \|(Ve_j)(e_j^T V^{-1})\|/|\lambda - D_{jj}|. \tag{2.22}$$

Then multiplying by $|\lambda - D_{ii}|/\|(Ve_i)(e_i^T V^{-1})\|$ gives

$$\frac{|\lambda - D_{ii}|}{\|(Ve_i)(e_i^T V^{-1})\|} \leq \|E\| \sum_{j=1}^{N} 1 = n\|E\|. \tag{2.23}$$

Multiplying by $\|(Ve_i)(e_i^T V^{-1})\|$ gives

$$|\lambda - D_{ii}| \leq n\|E\| \, \|(Ve_i)(e_i^T V^{-1})\|. \tag{2.24}$$

So far we are following the reasoning leading up to Theorem IV in [BF60] (also very similar to [Var04, Theorem 1.22]), which in slightly different notation states that all eigenvalues of B must lie in the subset of $\mathbb{C}$ satisfying (2.21), which is contained in the union $\bigcup_{i=1}^{n} M_i$, where M_i is the set satisfying condition (2.24).

In terms of matrix-valued functions, we have just derived inclusion regions for the eigenvalues of $T(z) = A - zI + E$ based on knowledge of the eigenvalues of A. This process carries over to the case where E is a matrix-valued function without much adjustment, leading to a generalization of [BF60, Theorem IV].

Theorem 2.3. *Let $T(z) = A - zI + E(z)$, where $E : \Omega \to \mathbb{C}^{n\times n}$ and $A \in \mathbb{C}$ with $AV = VD$. Then*

$$\Lambda(T) \subset M := \left\{ z \in \Omega : 1 \leq \|E(z)\| \sum_{k=1}^{n} |z - d_{jj}|^{-1} \|(Ve_j)(e_j^T V^{-1})\| \right\} \subset \bigcup_{i=1}^{n} M_i, \qquad (2.25)$$

where

$$M_i = \left\{ z \in \Omega : |z - d_{ii}| \leq n\|E(z)\| \, \|(Ve_i)(e_i^T V^{-1}\| \right\}. \qquad (2.26)$$

If $\mathcal{U}$ is a bounded connected component of M or $\bigcup_{i=1}^{n} M_i$ with the closure of $\mathcal{U}$ in $\mathbb{C}$ contained completely in Ω, then $\mathcal{U}$ contains the same number of eigenvalues of T and of A (counting multiplicity).

Proof. Previous discussion shows that $\Lambda(T) \subset M \subset \bigcup_{i=1}^{n} M_i$. The counting result follows the same way as in Theorem 2.1. □

Instead of using a bound on the norm of $E(z)$, we can use entrywise bounds. The next theorem is adapted from [BH13, Theorem 5.3]. The counting result therein is particularly convenient.

Theorem 2.4. *If $T = A - zI + E(z)$ as in the previous theorem, and if $|E(z)| \leq F$ entrywise on Ω, then $\Lambda(T) \subset \bigcup_{i=1}^{n} B_i$ where*

$$B_i = \{z \in \mathbb{C} : |z - d_{ii}| \leq \phi_i\}, \qquad \phi_i := n\|F\|_2 \sec(\theta_i) \qquad (2.27)$$

is the i-th Bauer-Fike disk and θ_i is the angle between Ve_i and $V^{-*}e_i$. Furthermore, if $\mathcal{U}$ is a bounded connected component of the union $\bigcup_{i=1}^{n} B_i$ consisting of exactly m Bauer-Fike disks, and if the closure of $\mathcal{U}$ in $\mathbb{C}$ is contained in Ω, then $\mathcal{U}$ contains exactly m eigenvalues of T.

Proof. We start from the point in the previous proof where we have deduced $(\lambda I - A)u = E(z)u$ for λ an eigenvalue of T, and take a different path. Premultiplying by V^{-1} gives $(\lambda I - D)x = V^{-1}E(z)Vx$ where $x = V^{-1}u$. The i-th row of this equation reads

$$(\lambda - D_{ii})x_i = \sum_{k=1}^{n}(V^{-1}EV)_{ik}x_k = \sum_{k=1}^{n}\left(\sum_{\ell=1}^{n}(V^{-1})_{i\ell}(EV)_{\ell k}\right)x_k \tag{2.28}$$

$$= \sum_{k=1}^{n}\left(\sum_{\ell=1}^{n}(V^{-1})_{i\ell}\left(\sum_{m=1}^{n}E_{\ell m}V_{mk}\right)\right)x_k \tag{2.29}$$

so

$$|\lambda - D_{ii}|\,|x_i| \leq \sum_{k=1e}^{n}\left(\sum_{\ell=1}^{n}|(V^{-1})_{i\ell}|\left(\sum_{m=1}^{n}|E_{\ell m}|\,|V_{mk}|\right)\right)|x_k| = \sum_{k=1}^{n}\left(|V^{-1}|\,|E|\,|V|\right)_{ik}\,|x_k| \tag{2.30}$$

$$= \sum_{k=1}^{n} e_i^T|V^{-1}|\,|E|\,|V|e_k\,|x_k| \tag{2.31}$$

where for a matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$, $|A|$ is defined as $|A| = (|a_{ij}|) \in \mathbb{R}^{n\times n}_{\geq 0}$.

Using the assumption on E, this becomes $|\lambda - D_{ii}|\,|x_i| \leq \sum_{k=1}^{n} e_i^T|V^{-1}|F|V|e_k\,|x_k|$. If we let $|x_i| = \max_k |x_k|$, then $|\lambda - D_{ii}| \leq \sum_{k=1}^{n} e_i^T|V^{-1}|\,F\,|V|e_k$. Since $\sum_{k=1}^{n}|V|e_k = |V|e$ where e is the vector of all ones, we have $|\lambda - D_{ii}| \leq e_i^T|V^{-1}|\,F\,|V|e$. Taking 2-norms, it follows that $|\lambda - D_{ii}| \leq \|e_i^T|V^{-1}|\,\|_2\,\|F\|_2\,\|\,|V|e\|_2$.

If we assume that the columns of V have Euclidean length 1, which we may do without loss of generality, then every entry of V is at most 1 in magnitude. Therefore $\|\,|V|e\|_2 \leq n$. Since $(V^{-*}e_i)^*(Ve_i) = e_i^TV^{-1}Ve_i = 1$, $\cos(\theta_i) = (\|V^{-*}e_i\|_2\,\|Ve_i\|_2)^{-1} = \|e_i^TV^{-1}\|_2$. Therefore $|\lambda - D_{ii}| \leq n\|F\|_2\sec(\theta_i)$, as desired.

The counting result follows the same way as in Theorem 2.1. □

2.3 Pseudospectral localization theorems

Recall that in Section 1.2 we defined the ε-pseudospectrum of a matrix-valued function $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$ as the set $\{z \in \Omega : \|T(z)^{-1}\| > \varepsilon^{-1}\}$, denoted by $\Lambda_\varepsilon(T)$, where $\|\cdot\|$ may be any norm induced by a vector norm. The equivalences mentioned in that definition are proved in the next proposition which parallels [TE05, Theorem 2.1].

Proposition 2.1. *Let $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$ be a matrix-valued function, and let $\varepsilon > 0$ be fixed. Define $\mathcal{E}$ as the set of matrix-valued functions $E : \Omega \to \mathbb{C}^{n\times n}$ satisfying $\|E(z)\| < \varepsilon$ on all of Ω, and define $\mathcal{E}_0$ to be the subset of $n \times n$ matrices E_0 with $\|E_0\| < \varepsilon$. Then the following definitions are equivalent:*

$$\Lambda_\epsilon(T) = \{z \in \Omega : \|T(z)^{-1}\| > \epsilon^{-1}\} \tag{2.32}$$

$$= \bigcup_{E\in\mathcal{E}} \Lambda(T + E) \tag{2.33}$$

$$= \bigcup_{E_0\in\mathcal{E}_0} \Lambda(T + E_0). \tag{2.34}$$

Proof. Denote the sets in (2.32), (2.33), and (2.34) as $\Lambda^1_\epsilon(T)$, $\Lambda^2_\epsilon(T)$, and $\Lambda^3_\epsilon(T)$. We break the proof into three steps:

$z \in \Lambda^2_\epsilon(T) \iff z \in \Lambda^3_\epsilon(T)$: If $T(z) + E(z)$ is singular for some $E \in \mathcal{E}$, then $T(z) + E_0$ is singular for $E_0 = E(z)$. Since $E_0 \in \mathcal{E}_0$, it follows that $z \in \Lambda^3_\epsilon(T)$. Conversely, if $T(z) + E_0$ is singular for some $E_0 \in \mathcal{E}_0$, then $T(z) + E(z)$ is singular for E the constant function E_0.

$z \notin \Lambda^1_\epsilon(T) \implies z \notin \Lambda^3_\epsilon(T)$: Suppose $\|T(z)^{-1}\| \le \epsilon^{-1}$. Then for any E_0 such that $\|E_0\| < \epsilon$, we have that $\|T(z)^{-1}E_0\| < 1$, so there is a convergent Neumann series[4] for $I + T(z)^{-1}E_0$. Thus, $(T(z) + E_0)^{-1} = (I + T(z)^{-1}E_0)^{-1}T(z)^{-1}$ is well defined.

[4] A Neumann series is the matrix version of a geometric series.

$z \in \Lambda^1_\epsilon(T) \implies z \in \Lambda^3_\epsilon(T)$: Eigenvalues of T belong to both sets, so we need only consider $z \in \Lambda^1_\epsilon(T)$ not an eigenvalue. So suppose $T(z)$ is invertible and $s^{-1} = \|T(z)^{-1}\| > \epsilon^{-1}$. Then $T(z)^{-1}u = s^{-1}v$ for some vectors u and v with unit norm; alternately, write $su = T(z)v$. Let $E_0 = -suw^*$, where w^* is a dual vector of v. Then $\|E_0\| = s < \epsilon$, and $T(z) + E$ is singular with v as a null vector. □

Remark 2.5. If $\|\cdot\| = \|\cdot\|_2$, then the 2-norm ε-pseudospectrum of T is the set $\Lambda_\varepsilon(T) = \{z \in \Omega : \sigma_{\min}(T(z)) < \varepsilon\}$.

The following result is nearly identical to the analogous statement for pseudospectra of a matrix [TE05, Theorem 4.2].

Proposition 2.2. *Suppose $T : \Omega \to \mathbb{C}^{n\times n}$ is analytic and $\mathcal{U}$ is a bounded connected component of $\Lambda_\epsilon(T)$ with the closure of U in $\mathbb{C}$ contained completely in Ω. Then $\mathcal{U}$ contains an eigenvalue of T.*

Proof. If $T(z)^{-1}$ is analytic on $\bar{\mathcal{U}}$, then by Lemma 2.4 $\|T(z)^{-1}\|$ is subharmonic on $\bar{\mathcal{U}}$. Therefore, the maximum of $\|T(z)^{-1}\|$ must be attained on the boundary. But $\|T(z)^{-1}\| = \epsilon^{-1}$ for $z \in \partial\mathcal{U}$, and $\|T(z)^{-1}\| > \epsilon^{-1}$ for $z \in \mathcal{U}$. Therefore, $T(z)^{-1}$ cannot be analytic on $\mathcal{U}$, i.e., there is an eigenvalue in $\mathcal{U}$. □

Keep in mind that we are interested in computing approximate eigenvalues for a matrix-valued function T by finding some $\hat{T} \approx T$ such that the eigenvalues of $\hat{T}$ are easy to compute. To see the relationship between pseudospectra of T and the eigenvalues of $\hat{T}$, consider $\hat{T} = T - E$ with $\|E(z)\| < \varepsilon$. Then $\hat{T}(z)^{-1} = (T(z) - E(z))^{-1} = (I - T(z)^{-1}E(z))^{-1}T(z)^{-1}$. If $\|T(z)^{-1}E(z)\| < 1$, then the Neumann series $\sum_{j=0}^{\infty}(T(z)^{-1}E(z))^j$ for $(I - T(z)^{-1}E(z))^{-1}$ converges; if $\|T(z)^{-1}\| \le \varepsilon^{-1}$ this is guaranteed. Therefore, if $\|T(z)^{-1}\| \le \varepsilon^{-1}$, then $\hat{T}(z)^{-1}$ is finite. Therefore if z

simultaneously satisfies $\|E(z)\| < \varepsilon$ and $\|T(z)^{-1}\| \leq \varepsilon^{-1}$, then z is not an eigenvalue of $\hat{T}$. Corresponding to this nonsingularity test is an inclusion result.

Lemma 2.5. *If λ is an eigenvalue of $\hat{T} = T - E$ satisfying $\|E(\lambda)\| < \varepsilon$, then $\|T(\lambda)^{-1}\| > \varepsilon^{-1}$, i.e., λ is in the ε-pseudospectrum of T.*

The nonsingularity test also gives us an opportunity to apply Lemma 2.1 to get a counting result.

Lemma 2.6. *If $\mathcal{U}$ is a bounded connected component of $\Lambda_\varepsilon(T)$ such that the closure of $\mathcal{U}$ in $\mathbb{C}$ is contained in the subset $\Omega_\varepsilon = \{z \in \Omega : \|E(z)\| < \varepsilon\}$, then $\mathcal{U}$ contains the same number of eigenvalues of T and $\hat{T} = T - E$.*

Proof. Notice that $\mathcal{U}$, as a connected component of $\Lambda_\varepsilon(T) = \{z \in \Omega : \|T(z)^{-1}\| > \varepsilon^{-1}\}$, is open in Ω. By assumption on $\mathcal{U}$, it follows that $\partial\mathcal{U}$ is completely contained in Ω_ε. Since $\|T(z)^{-1}\| = \varepsilon$ on $\partial\mathcal{U}$ by definition of $\Lambda_\varepsilon(T)$, $\partial\mathcal{U}$ is not in $\Lambda_\varepsilon(T)$. By the contrapositive of Lemma 2.5, no point of $\partial\mathcal{U}$ is an eigenvalue of $\hat{T} = T - E$. Furthermore, since $\|sE(z)\| \leq \|E(z)\|$ for $s \in [0, 1]$, no point of $\partial\mathcal{U}$ is an eigenvalue of $T - sE$ for any $s \in [0, 1]$. Therefore we can apply Lemma 2.1 with $\Gamma = \partial\mathcal{U}$ and the claim follows. □

The lemmas above result in the following theorem adapted from [BH13, Theorem 4.4].

Theorem 2.5. *Suppose $T, \hat{T} : \Omega \to \mathbb{C}^{n\times n}$ are regular analytic, define $E(z) = T(z) - \hat{T}(z)$ and let $\Omega_\varepsilon = \{z \in \Omega : \|E(z)\| < \varepsilon\}$. Then*

$$\Lambda(T) \cap \Omega_\varepsilon \subset \Lambda_\varepsilon(\hat{T}) \qquad \text{and} \qquad \Lambda(\hat{T}) \cap \Omega_\varepsilon \subset \Lambda_\varepsilon(T). \tag{2.35}$$

Furthermore, if $\mathcal{U}$ is a bounded connected component of $\Lambda_\varepsilon(T)$ such that the closure of $\mathcal{U}$ in $\mathbb{C}$ is in Ω_ε, then $\mathcal{U}$ contains exactly the same number of eigenvalues of T and $\hat{T}$ (counting multiplicity).

Proof. The second inclusion in (2.35) is Lemma 2.5 applied to every eigenvalue of T in Ω_ε, while the first follows the same way after interchanging the roles of T and $\hat{T}$. The counting result is a restatement of Lemma 2.6. □

2.3.1 Recommendations

Our intention is for Theorem 2.5 to be used to localize the eigenvalues of a matrix-valued function T in the following way:

1. Choose a region, call it U, where eigenvalues of T are desired.
2. Plot pseudospectra for T on U.
3. Form a matrix-valued function $\hat{T}$ such that the eigenvalues of $\hat{T}$ are easy to compute and $\|T - \hat{T}\|$ is small enough on U that Theorem 2.5 applies to some (ideally all) components of $\Lambda_\varepsilon(T)$ within U.
4. Compute the eigenvalues of $\hat{T}$ and use them in combination with the pseudospectra for T to deduce localization regions and counts for eigenvalues of T in U according to Theorem 2.5.
5. Compute the eigenvalues of T using an integral-based or iterative method.

Step 1 is up to the user. Step 2 can be accomplished by computing the values of $\|T(z)^{-1}\|^{-1}$ on a mesh of U and inputting the results into a good contour plotter, e.g. `contour` in MATLAB. This is straightforward in theory, but for large

problems this may be quite expensive. If it so happens that $\|\cdot\| = \|\cdot\|_2$ is the 2-norm (a choice we make in every pseudospectral plot in this thesis), then we have an important simplification according to the remark after Proposition 2.1, namely that $\|T(z)^{-1}\|_2^{-1} = \sigma_{\min}(T(z))$, and hence the boundary of the 2-norm ε-pseudospectrum of T is the set of points where $\sigma_{\min}(T(z)) = \varepsilon$. Then a contour plot should be created by computing the values of $\sigma_{\min}(T(z))$ on a mesh of U and inputting the results into a contour plotter. Either way, the resulting picture should show (parts of) the boundary of $\Lambda_\varepsilon(T)$ for several ε as contours, each labeled with their respective ε, and some of these contours will be closed.

Now we come to Step 3, which is the heart of the process. Suppose for concreteness that during Step 2 we found that μ is a value such that the boundary of the μ-pseudospectrum of T inside U consists of several closed contours. Then a reasonable next step is to find another matrix-valued function $\hat{T}$ such that $\|T - \hat{T}\| < \mu$ on U, or at least on a neighborhood of at least one component of $U \cap \overline{\Lambda_\mu(T)}$, and such that the eigenvalues of $\hat{T}$ are easy to compute. Many options immediately come to mind, such as taking $\hat{T}$ to be a Taylor polynomial that approximates T accurately near a certain point, or letting $\hat{T}$ be a Chebyshev polynomial that approximates T accurately near a certain interval (see [BH13]). A flexible approach to constructing such a $\hat{T}$ is to design it to be a rational matrix-valued function. For instance, supposing that T is analytic on and inside a curve Γ, we can use the Cauchy integral formula in combination with a quadrature

rule such as the trapezoid rule to do this:

$$T(z) = \frac{1}{2\pi i}\int_\Gamma \frac{T(\zeta)}{\zeta - z}\,dz \tag{2.36}$$

$$= \frac{1}{2\pi i}\int_0^1 \frac{T(\varphi(x))}{\varphi(x) - z}\,\varphi'(x)\,dx \qquad (\varphi : [0,1] \to \Gamma) \tag{2.37}$$

$$\approx \frac{1}{2\pi i}\sum_{j=1}^{N} \frac{T(\varphi(x_j))}{\varphi(x_j) - z}\,\varphi'(x_j)\,(x_j - x_{j-1}) \qquad (x_j = j/N) \tag{2.38}$$

$$= \sum_{j=1}^{N} \frac{W_j}{z_j - z} \tag{2.39}$$

where $z_j = \varphi(x_j)$ and $W_j = \frac{1}{2\pi i}T(z_j)\varphi'(x_j)(x_j - x_{j-1})$. By varying Γ and N, the region where $\|T(z) - \sum_j W_j/(z_j - z)\| < \mu$ can be manipulated. Once satisfactory Γ and N have been found, define $\hat{T}(z) = \sum_{j=1}^{N} W_j/(z_j - z)$ with these parameters. As a rational matrix-valued function, the eigenvalues of $\hat{T}$ can be computed via its linearization

$$\begin{bmatrix} 0 & W_1 & \dots & W_N \\ -I & z_1 I & & \\ \vdots & & \ddots & \\ -I & & & z_N I \end{bmatrix} - z \begin{bmatrix} 0 & & & \\ & -I & & \\ & & \ddots & \\ & & & -I \end{bmatrix} \tag{2.40}$$

whose Schur complement [Zha05] equals $\hat{T}$. Obviously this method of constructing a rational approximation is out of the question for large problems, but in practice usually only a few entries of a genuinely nonlinear matrix-valued function need to be approximated and we can use a similar process to accomplish the same thing while using less space. This is done in Chapter 5 for a matrix-valued function that comes from discretizing a linear differential equation with nonlinearity in the boundary condition. Also, sometimes rational approximations naturally appear in a series expansion, which we take advantage of in our treatment of the `gun` problem in Chapter 3.

Step 4 is now simple. Suppose we have decided on a certain $\hat{T}$ in Step 3,

where $\hat{T} \approx T$. Then we compute and plot the eigenvalues of $\hat{T}$ in the region of interest U. Suppose now that $\mathcal{V}$ is a component of $\Lambda_\mu(T)$ such that $\|\hat{T} - T\| < \mu$ on $\bar{\mathcal{V}}$. By Theorem 2.5, $\mathcal{V}$ must contain the same number of eigenvalues of $\hat{T}$ and T.[5] Since $\Lambda(T) \subset \Lambda_\varepsilon(T)$, this method allows us to localize all eigenvalues of T within $\Omega_\mu = \{z \in \Omega : \|T - \hat{T}\| < \mu\}$.

Lastly, we come to Step 5. Assume that all eigenvalues of T in U have been localized, and let $\mathcal{V}$ be a bounded component of $\Lambda_\mu(T)$ in U. Suppose m eigenvalues of $\hat{T}$ are in $\mathcal{V}$, so m eigenvalues of T must be also. Then there are (at least) three standard approaches to computing those m eigenvalues of T that use the information provided by the eigenvalues of $\hat{T}$.[6] First, we can use a contour integral-based method such as [VBK16], [Bey12], or [AST+09] to compute the m eigenvalues of T in $\mathcal{V}$. With these methods, it is important to choose an integration contour that is not too near any eigenvalues, i.e., $\|T(z)^{-1}\|$ must not be too large. Choosing a contour that surrounds $\mathcal{V}$ but is not too close to any other component of $\Lambda_\mu(T)$ will eliminate this issue. In addition, knowing the number of eigenvalues within $\mathcal{V}$ in advance simplifies the choice of parameters for the algorithms and saves on execution time. Second, an iterative, Newton-based method method such as [Kre09] can be used for simultaneous computation of the m eigenvalues of T in $\mathcal{V}$. Such an algorithm requires a set of initial guesses, which we could provide with eigenvalues of $\hat{T}$ or with their refinements. We can then validate the results by checking whether all computed eigenvalues are still within $\mathcal{V}$. This course is most appropriate if the component $\mathcal{V}$ is small or if we have other reasons to believe that the eigenvalues of $\hat{T}$ are very close to those of T within $\mathcal{V}$. Third, we could use a bordered Newton iteration [Gov00,

[5] Of course, $\mathcal{V}$ contains at least one eigenvalue of T (see Proposition 2.2), so $\mathcal{V}$ must contain at least one eigenvalue of $\hat{T}$ as well.

[6] Each will be demonstrated on at least one example in Chapter 3.

Ch. 3] on each eigenvalue of $\hat{T}$ in $\mathcal{V}$ individually and hope for convergence to an eigenvalue of T. If the limits are distinct, all in $\mathcal{V}$, and there are m of them, we have found the m eigenvalues of T in $\mathcal{V}$.

CHAPTER 3

GALLERY OF EXAMPLES

3.1 Naïve application of nonlinear Gershgorin theorem

In this section, we present matrix-valued functions $T(z)$ from the literature where a naïve splitting into diagonal and off-diagonal parts give useful localization regions.

3.1.1 Single delay PDDE I

This example is taken from the example [Eff13, §5.1] where eigenvalues of the partial delay differential equation

$$u_t(x,t) = u_{xx}(x,t) + a_0 u(x,t) + a_1(x)u(x,t-\tau), \qquad u(0,t) = u(\pi,t) = 0 \tag{3.1}$$

$$a_0 = 20, \quad a_1(x) = -4.1 + x(1-\exp(x-\pi)), \quad \tau = 0.2 \tag{3.2}$$

are studied. Following [Eff13], discretizing with the centered difference approximation $u_{xx}(x,t) \approx (u(x+h,t) - 2u(x,t) + u(x-h,t))/h^2$ at grid points $x_j = jh$, $h = \pi/(n+1)$, $n = 1000$, $j = 1,...,n$ gives

$$\begin{bmatrix} u_{xx}(x_1,t) \\ \vdots \\ u_{xx}(x_n,t) \end{bmatrix} = \frac{1}{h^2}\left(\begin{bmatrix} u(x_2,t) \\ \vdots \\ u(x_{n+1},t) \end{bmatrix} - 2\begin{bmatrix} u(x_1,t) \\ \vdots \\ u(x_n,t) \end{bmatrix} + \begin{bmatrix} u(x_0,t) \\ \vdots \\ u(x_{n-1},t) \end{bmatrix}\right). \tag{3.3}$$

By the boundary conditions, $u(x_0, t) = 0$ and $u(x_{n+1}, t) = 0$. Therefore, this discretization of (3.1) yields[1]

$$\begin{bmatrix} u_t(x_1,t) \\ \vdots \\ u_t(x_n,t) \end{bmatrix} = \underbrace{\left(\frac{1}{h^2} \begin{bmatrix} -2 & 1 & \\ 1 & \ddots & 1 \\ & 1 & -2 \end{bmatrix} + a_0 I \right)}_{A_0} \begin{bmatrix} u(x_1,t) \\ \vdots \\ u(x_n,t) \end{bmatrix} + \underbrace{\begin{bmatrix} a_1(x_1) & & \\ & \ddots & \\ & & a_1(x_n) \end{bmatrix}}_{A_1} \begin{bmatrix} u(x_1,t-\tau) \\ \vdots \\ u(x_n,t-\tau) \end{bmatrix}. \tag{3.4}$$

Defining $v(t) = [u(x_1, t), ...u(x_n, t)]^T$ gives the first order delay differential equation $\dot{v}(t) = A_0 v(t) + A_1 v(t - \tau)$. The asymptotic growth and decay of its solutions are determined by the eigenvalues of $T(z) = zI - A_0 - A_1 \exp(-\tau z)$.

Since the off-diagonal part of $T(z)$ is contained in the constant term, we expect that the diagonal of $T(z)$ becomes more and more dominant as $|z| \to \infty$. Motivated by this, we apply Theorem 2.1 to $T(z)$. The resulting inclusion region is unbounded and symmetric about the real axis. Figure 3.1 (left) shows the part of the inclusion region in $[-65, 30] \times [0, 32000] \subset \mathbb{C}$. By the time z gets to the top of this rectangle, the inclusion region has split into connection components (see Figure 3.1 (center)), each of which must contain the same number of eigenvalues of T and zeros of the diagonal entries of T (see the counting result in Theorem 2.1). The latter are computed using `lambertw` in MATLAB and plotted. See [CGH+96] for more on the Lambert W function. Near to the origin the inclusion region is far less useful (see Figure 3.1 (right)), being comb-shaped rather than a union of several components. We will return to this example in Section 3.2 to present very tight inclusion regions derived another way.

[1]Note the error in the definition of A_0 in [Eff13, §5.1] where $h^2 = \left(\frac{\pi}{n+1}\right)^2$ rather than $1/h^2$ is used.

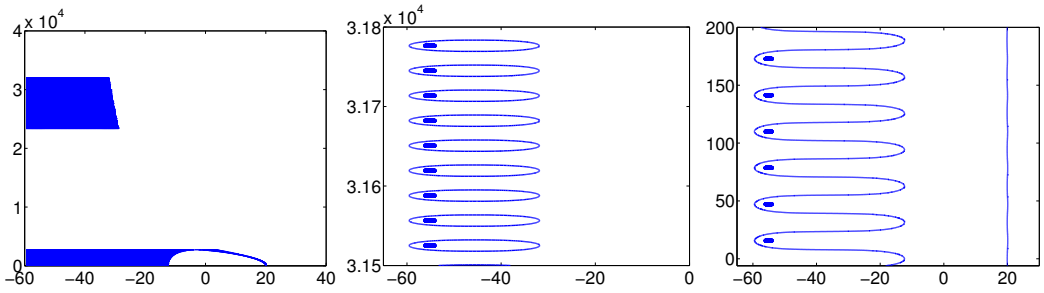

Figure 3.1: Naïve inclusion regions for a single delay PDDE.

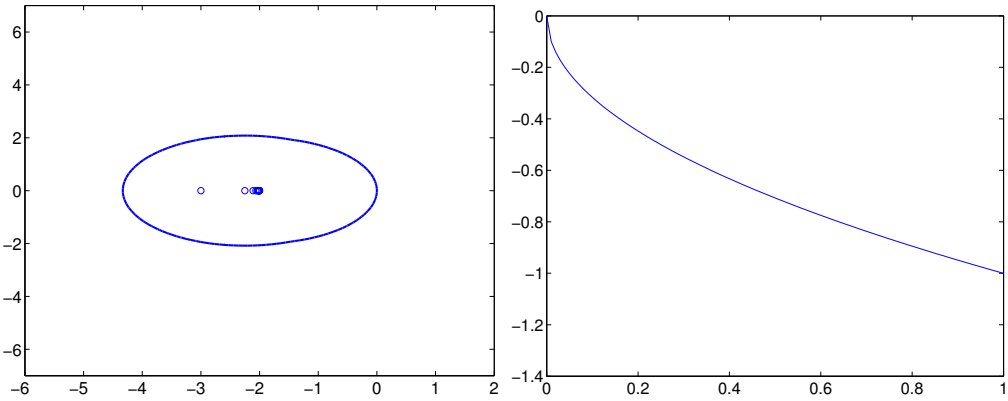

Figure 3.2: Naïve inclusion region for `fiber` problem and graph of $s(z)$.

3.1.2 Fiber

The `fiber` problem from [BHM+13], [XH10] is of the form

$$F(z) = A - zI + s(z)e_n e_n^T \tag{3.5}$$

where $n = 2400$, e_n is the last column of the $n \times n$ identity, A is real and tridiagonal, and $s(z)$ involves a ratio of Bessel functions taking multiples of $\sqrt{z}$ as arguments.

With the splitting $F(z) = \hat{F}(z) + E(z)$ into diagonal and off-diagonal parts, Theorem 2.1 gives the Gershgorin region inside the thick blue line shown in Figure 3.2 (left), where the eigenvalues of $\hat{F}$ are shown as well ($\circ$). Unfortunately, this contour encircles the origin, which is a branch point for s. Therefore we cannot use the counting result from Theorem 2.1, and this means we cannot be sure whether we have computed all the eigenvalues of F using some iterative

algorithm. Integral-based algorithms cannot be used effectively here because F is not analytic on any contour surrounding the inclusion region.

Luckily, according to [BHM+13] and the sources cited therein, only the real and positive eigenvalues are needed. It is known (see [XH10, Example 3] and Figure 3.2 (right)) that s is negative and decreasing on $(0, \infty)$. This allows us to analytically derive an interval in which real, positive eigenvalues must lie. From Theorem 2.1, the Gershgorin regions associated to each row are in this case $G_j = \{|A_{jj} - z| \leq r_j\}$, $j = 1, ..., n-1$, and $G_n = \{|A_{nn} - z + s(z)| \leq r_n\}$. Since A is tridiagonal, $r_1 = |A_{12}|$, $r_n = |A_{n,n-1}|$, and $r_j = |A_{j,j-1}| + |A_{j,j+1}|$, $j = 2, ..., n-1$. Then G_j is circular for $j \leq n-1$ and G_n is of some unknown shape, possibly consisting of multiple components. Recalling that A is real, one can compute how far to the right each of the G_1, ..., G_{n-1} reach as $\max_{j \leq n-1} \left(A_{jj} + r_j\right) \approx 3.5 \times 10^{-5}$. As for G_n, the fact that all three of A_{nn}, $s(z)$, and $-z$ are negative for $z > 0$ implies that $|A_{nn} - z + s(z)| = |A_{nn}| + |z| + |s(z)|$ for $z > 0$. Therefore, $z \in G_n \cap (0, \infty)$ is equivalent to $|z| + |s(z)| \leq |A_{n,n-1}| - |A_{nn}|$, where the right-hand side is $\approx 2.0818 \times 10^{-4}$. Clearly this implies $|z| \leq 2.0818 \times 10^{-4}$.

Hence we have derived that any positive, real eigenvalue of F must lie in the interval $J = (0, 2.0818 \times 10^{-4}]$. Using 8 steps of bordered Newton iteration [Gov00, Ch. 3] on T with the right endpoint of J as an initial guess produces

$$\lambda = 7.13949430588745 \times 10^{-7}, \qquad \sigma_{\min} T(\lambda) = 2.1 \times 10^{-16}. \tag{3.6}$$

The table in [XH10, p. 233] shows agreement with this result.

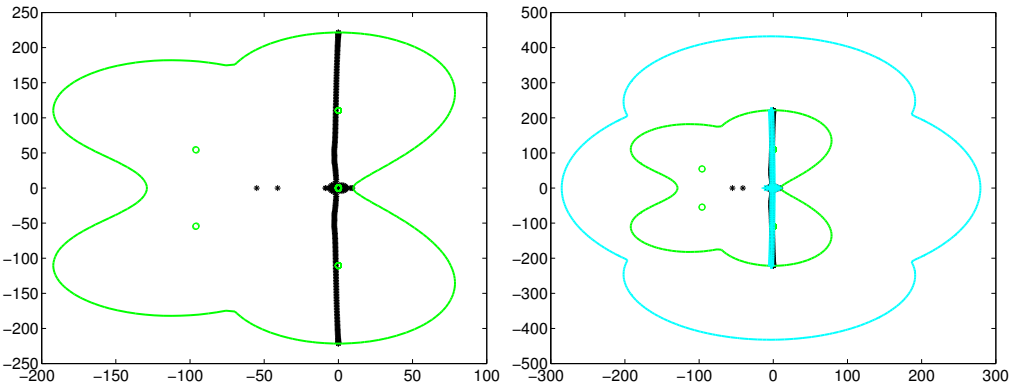

Figure 3.3: Inclusion regions for `planar_waveguide` problem.

3.1.3 Planar waveguide

The `planar_waveguide` problem from [BHM+13] is of the form $T(z) = A_4z^4 + A_3z^3 + A_2z^2 + A_1z + A_0$, where each A_j is in $\mathbb{R}^{129\times129}$. Both A_1 and A_3 are diagonal and rank 2, and the rest are tridiagonal, full rank, and have no zeros on their diagonals. Additionally, the magnitudes of the diagonal entries of A_0, A_2, and A_4 exceed the magnitudes of the off-diagonal entries by about a factor of 4. This suggests we can get reasonable inclusion regions by applying Theorem 2.1 without needing to do any transformations. Accordingly, we use the splitting $T(z) = \hat{T}(z) + E(z)$ where $\hat{T}(z)$ is the diagonal part of $T(z)$. The resulting inclusion region for the eigenvalues of T, and the eigenvalues ($\circ$) of $\hat{T}$, are pictured in Figure 3.3 (left). The eigenvalues of T ($*$), computed using MATLAB's `polyeig` are plotted as well.

The matrix A_4 is symmetric positive definite, and A_0 is a multiple of A_4, say $A_0 = cA_4$. Therefore a Cholesky decomposition $A_4 = R^TR$ can be used for an eigenvalue-preserving transformation of T: $R^{-T}T(z)R^{-1} = Iz^4+B_3z^3+B_2z^2+B_1z+cI$, where $B_j = R^{-T}A_jR^{-1}$. At this point, we can choose to diagonalize any one of B_3, B_2, or B_1. After experimenting, the best choice of the three is to diagonalize B_2,

since it is full rank (unlike B_3) and large in norm. With $B_2V = VD$, we transform T again to become

$$\tilde{T}(z) = V^{-1}R^{-T}T(z)R^{-1}V = Iz^4 + C_3z^3 + Dz^2 + C_1z + cI, \qquad C_j = V^{-1}B_jV.$$

Then we split $\tilde{T}(z) = \check{T}(z) + F(z)$ where $\check{T}(z)$ is the diagonal part of $\tilde{T}(z)$, and apply Theorem 2.1. Figure 3.3 (right) shows the resulting inclusion region and the eigenvalues of $\check{T}$ (+). Although the eigenvalues of $\check{T}$ approximate those of T much better than do the eigenvalues of $\hat{T}$, the inclusion region is worse. This shows that the simple splitting $T = \hat{T} + E$ sometimes gives better inclusion regions than a more sophisticated analysis does.

3.1.4 Butterfly I

The `butterfly` problem from [BHM+13] has nothing to do with butterflies except that its spectrum resembles one (see Figure 3.4 (left)). The matrix-valued function whose eigenvalues we want is a matrix polynomial

$$T(z) = A_4z^4 + A_3z^3 + A_2z^2 + A_1z + A_0 \tag{3.7}$$

where each $A_j \in \mathbb{R}^{64\times64}$ is either a symmetric or a skew-symmetric Kronecker product and each A_j is zero off diagonals -8, -1, 0, 1, and 8. If we do not exploit any of this structure and apply Theorem 2.1 according to the splitting of T into diagonal and off-diagonal parts, we obtain the inclusion region in Figure 3.4 (right). Unfortunately, the inclusion region we obtain in this manner is unbounded. We will revisit this example three times in this chapter, using strategies of diagonalization and block structure exploitation.

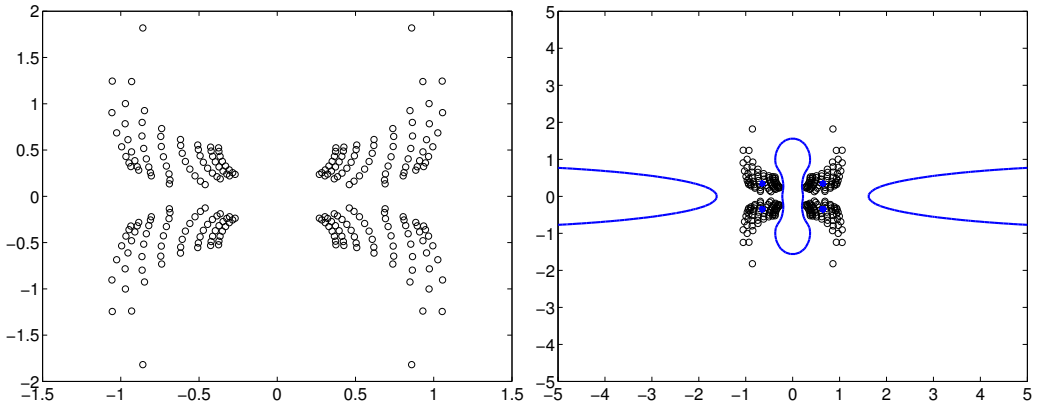

Figure 3.4: Spectrum and naïve inclusion regions for `butterfly` problem.

3.2 Diagonalizing the dominant term

In practice, a matrix-valued function is often the sum of a few terms, such as $T(z) = A_2z^2+A_1z+A_0$, or $T(z) = zI-A-Be^{-z}$. Diagonalizing one of the terms makes the diagonal of $T(z)$ bigger where the term is dominant. Applying Theorem 2.1 to the resultant transformed problem gives inclusion regions that are likely to be tighter in the region where the diagonalized term is dominant. In addition to the examples presented in this section, see the treatment of the `time_delay` problem from [BHM+13] in [BH13].

3.2.1 Single delay PDDE II

We return to the example presented in Section 3.1, of the form

$$T(z) = zI - A_0 - A_1 \exp(-\tau z) \tag{3.8}$$

where A_0 is tridiagonal and A_1 is diagonal. Recall that T comes from asymptotic stability analysis for solutions to the delay equation $\dot{v}(t) = A_0v(t) + A_1v(t-\tau)$.

Previously we computed inclusion regions which were useful very far from

the origin. However, the asymptotic stability of solutions to the delay equation hinges on whether the eigenvalues of T are all in the left half-plane, and growth of solutions is largely determined by several of the right-most eigenvalues, that is, the eigenvalues with largest real part. Examining Figure 3.1 makes it clear the eigenvalues near the origin are the ones of interest for asymptotic analysis, so we should concentrate on localizing those (compare with [§5.1][Eff13], also discussed in Section 1.1). Because A_0 comes from discretizing a differential equation and has $1/h^2$ dependence for tiny parameter h, the behavior of $T(z)$ is dominated by A_0 if $|z|$ is small.[2] Therefore we diagonalize A_0 to obtain the matrix-valued function

$$\tilde{T}(z) = zI - D - \tilde{A}_1 \exp(-\tau z), \qquad A_0 D = DV, \quad \tilde{A}_1 = V^{-1} A_1 V \tag{3.9}$$

with the same spectrum as T. Applying Theorem 2.1 to $\tilde{T}$ gives the inclusion regions shown in Figure 3.5 (left) on top of the inclusion regions derived in Section 3.1. The zeros of the diagonal entries of $\tilde{T}$ are also plotted ($\circ$). Notice that the inclusion region components around the three largest eigenvalues are so tight that the rightmost is not visible and the other two are barely so.

Because the part of the spectrum near the origin has been localized, we can now compute those eigenvalues and be confident about the results. We will use two methods: [Kre09, Algorithm 1] and [Bey12, Algorithm 1], which we call Kressner's algorithm and Beyn's algorithm, respectively. These algorithms both work for general, analytic matrix-valued functions. Kressner's algorithm is iterative and requires initial guesses, while Beyn's algorithm computes eigenvalues within a user-specified simple, closed contour.

[2] It is because of the dominance of A_0 near the origin, as well as the growth of the $\exp(-\tau z)$ term in the left half-plane, that the "envelope curve" inclusion region $|z| \leq \|A_0\|_2 + \|A_1\|_2 \exp(-\tau z)$ (see Section 1.3) is not helpful for this problem.

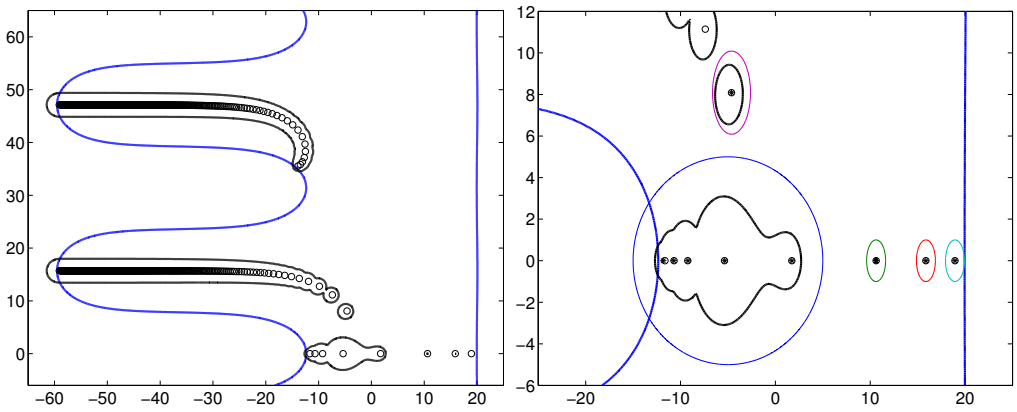

Figure 3.5: Diagonalization-based inclusion regions and eigenvalues for single delay PDDE example.

$\hat{\lambda}$	$\sigma_{\min}T(\hat{\lambda})$	λ	$\sigma_{\min}T(\lambda)$	$\lvert\lambda-\hat{\lambda}\rvert$
-4.622505e+00 +8.086895e+00i	7.87e-03	-4.620537e+00 +8.083313e+00i	3.88e-11	4.09e-03
+1.061861e+01 +0.000000e+00i	3.06e-05	+1.061857e+01 +0.000000e+00i	3.45e-11	3.31e-05
+1.586818e+01 +0.000000e+00i	6.14e-06	+1.586817e+01 +0.000000e+00i	9.19e-12	6.30e-06
+1.893224e+01 +0.000000e+00i	8.09e-06	+1.893225e+01 +0.000000e+00i	1.96e-11	8.20e-06
-1.163801e+01 +0.000000e+00i	8.94e-01	-1.181831e+01 +0.000000e+00i	4.99e-12	1.80e-01
-1.068312e+01 +0.000000e+00i	1.47e-01	-1.071767e+01 +0.000000e+00i	2.65e-11	3.46e-02
-9.237972e+00 +0.000000e+00i	6.73e-02	-9.215977e+00 +0.000000e+00i	2.99e-11	2.20e-02
-5.358304e+00 +0.000000e+00i	1.38e-02	-5.342532e+00 +0.000000e+00i	1.54e-11	1.58e-02
+1.735173e+00 +0.000000e+00i	8.20e-04	+1.733673e+00 +0.000000e+00i	5.96e-12	1.50e-03

Table 3.1: Error table for single delay PDDE I.

Specifically, we will compute the eigenvalues in the origin-containing component, the three rightmost eigenvalues, and the eigenvalue near $-5 + 8i$. We provide the eigenvalues of $\tilde{T}$ in those components as initial guesses for Kressner's algorithm applied to T. The initial guesses ($\circ$) and results ($*$) are plotted in Figure 3.5 (right) and tabulated in Table 3.1. The table columns are the initial guesses $\hat{\lambda}$, their residuals $\sigma_{\min}T(\hat{\lambda})$, the true eigenvalues λ of T computed with Kressner's algorithm, their residuals, and finally the absolute errors $|\lambda - \hat{\lambda}|$. The last show that the eigenvalues of the diagonal part of $\tilde{T}$ are already quite close to the eigenvalues of T.

The contours drawn around the relevant components in Figure 3.5 (right) were used as inputs to Beyn's algorithm, which is sensitive to the choice of

contour. In particular, if a contour is too close to an eigenvalue, Beyn's algorithm becomes inefficient or inaccurate [Bey12, Remarks 3.5(d)]. Because the components of the localization regions are well-separated, we were able to choose contours that avoided such issues. The pictured contours are an ellipse $-5 + 10\cos(t) + 5i\sin(t)$, unit-radius circles centered at each of the three rightmost eigenvalues of the diagonal part of $\tilde{T}$, and a radius 2 circle centered at the eigenvalue of the diagonal part of $\tilde{T}$ near $-5 + 8i$. The trapezoid rule parameter N was set to 1000 for the ellipse, 100 for the unit circles, and 200 for the circle of radius 2. The parameter ℓ was set to the number of eigenvalues in each contour, respectively, obtained through the counting result in Theorem 2.1 and the eigenvalues of the diagonal part of $\tilde{T}$. With the other algorithm parameters set to `tolrank` $= 10^{-8}$, `tolres` $= 10^{-6}$ and `maxcond` $= 10$, the eigenvalues computed by Beyn's algorithm agree with those computed by Kressner's algorithm to within 10^{-10} and have residuals with the same order of magnitude.

3.2.2 CD player

The `cd_player` problem from [BHM+13] is of the form $T(z) = Iz^2 + Cz + K$, where C and K are in $\mathbb{R}^{60\times 60}$. Since the term with largest growth is a multiple of the identity, we diagonalize the term with the next largest growth. By doing so, we expect that inclusion regions far from the origin will be tight. For $CV = VD$, we define $\tilde{T}(z) = Iz^2 + Dz + V^{-1}KV$, and split $\tilde{T}(z) = \hat{T}(z) + E(z)$ into diagonal and off-diagonal parts, respectively. Applying Theorem 2.1 with this splitting gives inclusion regions so tight that they can't be seen on a plot that includes all eigenvalues of $\hat{T}$ (which are estimates for eigenvalues of $\tilde{T}$ and therefore for T). Figure 3.6 shows this and some inclusion regions near a few particular

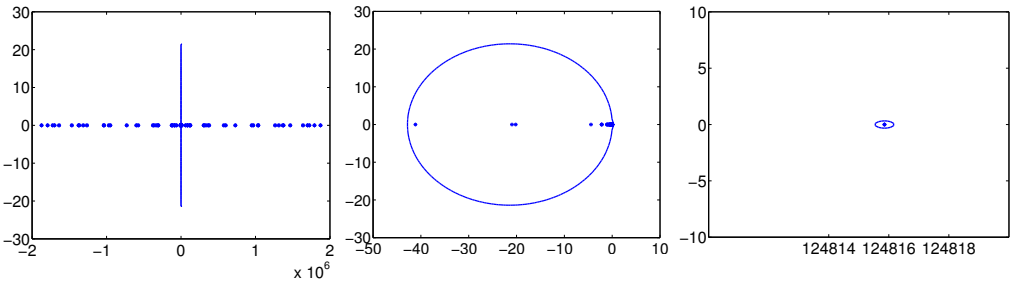

Figure 3.6: Inclusion regions for `cd_player` problem.

eigenvalues.

Since T is a matrix polynomial, we take this opportunity to compare with inclusion regions obtained through localization results tailored to the polynomial case. Specifically, from [BNS13, Theorem 2.1], the positive roots of

$$x^2 = \|C\|x + \|K\| \tag{3.10}$$

$$x = \|C^{-1}\|x^2 + \|C^{-1}K\| \tag{3.11}$$

$$1 = \|K^{-1}\|x^2 + \|K^{-1}C\| \tag{3.12}$$

will give us inclusion and exclusion annuli. For this example we use the 2-norm, and other choices of norm do not make a perceptible difference in the results.

Theorem 2.1.3 in [BNS13] states that there can be no eigenvalues of T with modulus less than the positive root of (3.12), which turns out to be $\approx 2.2 \times 10^{-4}$. In this context, this result is not very interesting because the exclusion disk we infer is so small. Theorem 2.1.4 in [BNS13] also states that no eigenvalues of T can have modulus greater than the positive root of (3.10), which is $\approx 1.07 \times 10^7$. This is quite a bit less useful than the inclusion regions we have derived, since our inclusion regions consist of a large central component and individual components around each eigenvalue with modulus greater than about 1000, and the latter components get tighter and tighter as modulus of the eigenvalue increases. Lastly, Theorem 2.1.1 states that no eigenvalues can have modulus

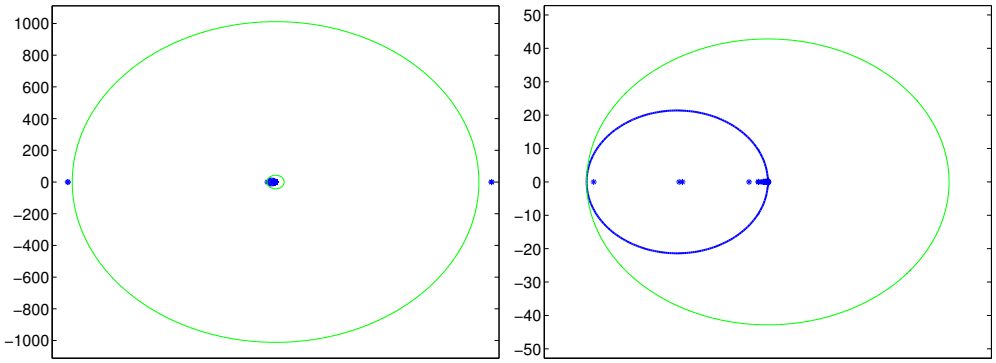

Figure 3.7: Comparison with matrix polynomial inclusion regions for the `cd_player` problem.

between the two positive roots of (3.11). The origin-centered circles with radii equal to these two positive roots are shown in Figure 3.7 (left) as green curves. The outer circle does not bound eigenvalue moduli from below as tightly as our inclusion regions, as discussed above. As for the inner circle, we can see from Figure 3.7 (right) that our inclusion region for the eigenvalues of small modulus is bounding moduli just as well, and it is significantly less generous because it is not forced to be axisymmetric.

3.2.3 Butterfly II

The `butterfly` problem from [BHM+13] is a matrix polynomial of the form $T(z) = \sum_{j=0}^{4} A_j z^j$, where each A_j is a Kronecker product. We have already approached the problem of localizing the eigenvalues of T once in Section 3.1. In this section, we will use the fact that the Kronecker products defining A_0, A_2, and A_4 are related and can be diagonalized simultaneously.

Specifically, M_0, M_1, and c are certain matrices used to define the matrices A_j in the original source [MW02] cited by [BHM+13]. If $M_0 V_0 = V_0 D_0$ is the

eigendecomposition of the real, symmetric matrix M_0, and we define $V = V_0 \times V_0$ and $X = V_0^* M_1 V_0$, then

$$B_0 := V^* A_0 V = c_{01}(I \otimes D_0) + c_{02}(D_0 \otimes I) \tag{3.13}$$

$$B_1 := V^* A_1 V = c_{11}(I \otimes X) + c_{12}(X \otimes I) \tag{3.14}$$

$$B_2 := V^* A_2 V = -6c_{21} I + 6c_{21}(I \otimes D_0) - 6c_{22} I + 6c_{22}(D_0 \otimes I) \tag{3.15}$$

$$B_3 := V^* A_3 V = c_{31}(I \otimes X) + c_{32}(X \otimes I) \tag{3.16}$$

$$B_4 := V^* A_4 V = 6c_{41} I - 6c_{41}(I \otimes D_0) + 6c_{42} I - 6c_{42}(D_0 \otimes I) \tag{3.17}$$

gives the transformed matrix-valued function

$$S(z) = B_4 z^4 + B_3 z^3 + B_2 z^2 + B_1 z + B_0 \tag{3.18}$$

with the same eigenvalues as T. Splitting $S(z) = \hat{S}(z) + E(z)$ into the sum of its diagonal and off-diagonal parts, respectively, and applying Theorem 2.1 gives localization regions shown in Figure 3.8 (left). Eigenvalues of $\hat{S}$ $(*)$ and the true eigenvalues of T $(\circ)$ are shown as well. For this particular problem, using the Kronecker product structure gives localization regions that are much better than the unbounded localization region that comes from applying Theorem 2.1 directly to T. The latter is shown in Figure 3.8 (right) in blue for comparison. However, the localization region we derive in this section is not as good as any of the ones we will derive by using a Cholesky factorization in Section 3.3 (see Figure 3.12). Therefore, for this particular problem, it is better to prioritize diagonalization of the terms with fastest growth than to diagonalize the lower order terms. The best option of all will be shown in Section 3.5.

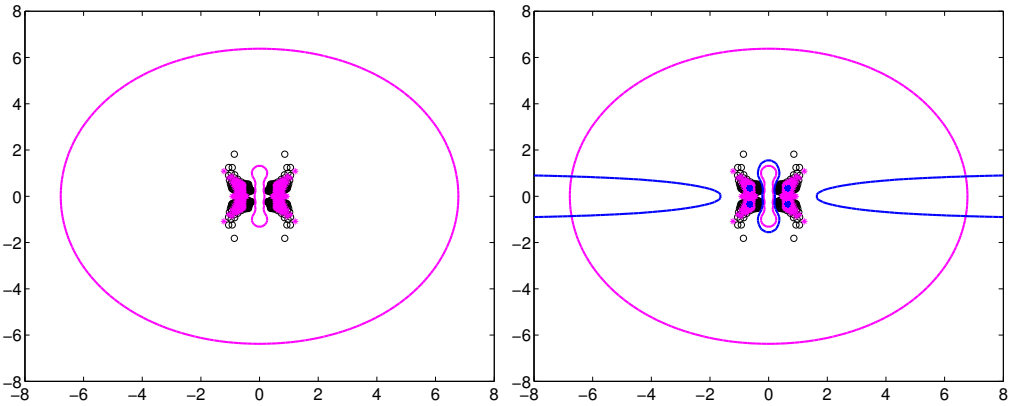

Figure 3.8: Localization region from simultaneous diagonalization of z^4, z^2, and constant terms in `butterfly` problem.

3.2.4 Hospital

The `hospital` problem from [BHM+13] is of the form $T(z) = z^2I + Cz + K$, where K and C are order 10^4 and 10 in 2-norm, respectively. To make inclusion regions from Theorem 2.1 as tight as possible we need to increase the size of the diagonal relative to the off-diagonal. Therefore we diagonalize K. If $KV = VD$, then an equivalent transformed problem is $S(z) = z^2I + \tilde{C}z + D$, where $\tilde{C} = V^{-1}CV$. Because the sums $\sum_{k\neq j} |\tilde{C}_{jk}|$ are less than 10^{-6} for all rows j, the components of the inclusion region obtained via Theorem 2.1 applied to S are so tight as to be invisible in Figure 3.9 (left). The actual eigenvalues of T ($\circ$) and the eigenvalues of $\hat{S}$ ($*$) are also plotted. Each component contains exactly one eigenvalue of $\hat{S}$, and by the counting result in Theorem 2.1, exactly one eigenvalue of T. Since the inclusion regions are so tight, we can analytically derive disks centered at each eigenvalue of $\hat{S}$ that contain the corresponding eigenvalue of T.

First, we recall that the Gershgorin region corresponding to the j-th row of S is $G_j = \{z : |z^2 + \tilde{C}_{jj}z + D_{jj}| \leq |z| \sum_{k\neq j} |\tilde{C}_{jk}|\}$. Factoring the quadratic into $(z - z_-^{(j)})(z - z_+^{(j)})$, we have $G_j = \{z : |z - z_-^{(j)}| \cdot |z - z_+^{(j)}| \leq |z| \sum_{k\neq j} |\tilde{C}_{jk}|\}$. Both $z_-^{(j)}$

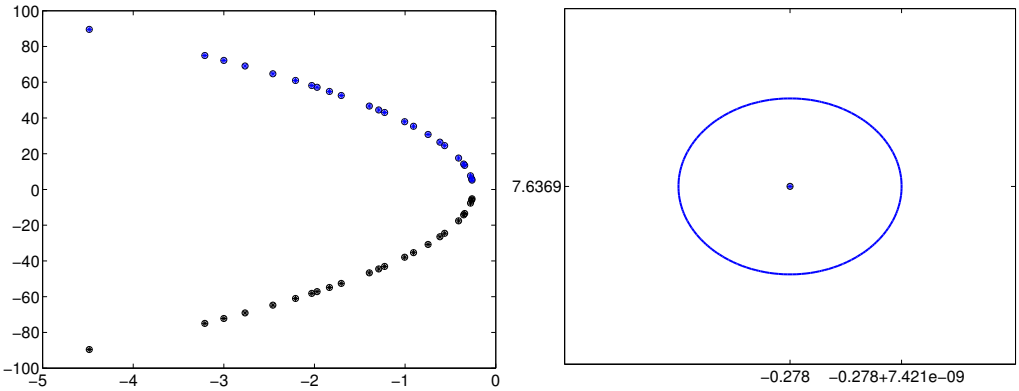

Figure 3.9: Eigenvalues and inclusion regions for the `hospital` problem.

and $z_+^{(j)}$ are complex conjugate eigenvalues of $\hat{S}$ due to the realness of $\tilde{C}$ and D. We would like to find a radius r_j such that the disk $B(z_+^{(j)}, r_j)$ of radius r_j centered at $z_+^{(j)}$ contains the component of G_j surrounding $z_+^{(j)}$ (and similarly for $z_-^{(j)}$). Using the fact that $\sum_{k\neq j} |\tilde{C}_{jk}| < 10^{-6}$, it's enough that $B(z_+^{(j)}, r_j)$ contain $\tilde{G}_j = \{z : |z - z_-^{(j)}| \cdot |z - z_+^{(j)}| \leq 10^{-6}|z|\}$.

Assume that $r_j < \frac{1}{2}|z_+^{(j)} - z_-^{(j)}|$, which ensures that the disks surrounding $z_+^{(j)}$ and $z_-^{(j)}$ do not intersect. We need r_j such that $|z - z_-^{(j)}| = r_j$ implies that $z \notin \operatorname{int} \tilde{G}_j$. The latter is equivalent to $|z - z_-^{(j)}| \cdot |z - z_+^{(j)}| \geq 10^{-6}|z|$. Now, if $|z - z_-^{(j)}| = r_j$, the assumption on r_j implies $|z - z_+^{(j)}| > \frac{1}{2}|z_+^{(j)} - z_-^{(j)}|$ and $|z| \leq r_j + |z_-^{(j)}|$. Therefore, if

$$r_j \frac{1}{2}|z_+^{(j)} - z_-^{(j)}| \geq 10^{-6}(r_j + |z_-^{(j)}|), \tag{3.19}$$

then $r_j|z - z_+^{(j)}| \geq 10^{-6}|z|$ (i.e., $|z - z_-^{(j)}| = r_j$ implies $z \notin \tilde{G}_j$, as desired). The condition (3.19) is equivalent to

$$r_j \geq \frac{10^{-6}|z_-^{(j)}|}{\frac{1}{2}|z_+^{(j)} - z_-^{(j)}| - 10^{-6}}. \tag{3.20}$$

the right-hand side is less than $\frac{1}{2}|z_+^{(j)} - z_-^{(j)}|$ for all j, so there is no inconsistency with the original assumption on r_j. Obviously we take r_j as small as possible,

namely

$$r_j = \frac{10^{-6}|z_-^{(j)}|}{\frac{1}{2}|z_+^{(j)} - z_-^{(j)}| - 10^{-6}}. \tag{3.21}$$

As can be seen in Table 3.2, each r_j is approximately 10^{-6}. Therefore the error in the eigenvalues of $\hat{S}$ in approximating the eigenvalues of S is less than or equal to 10^{-6}. The rest of the table shows the eigenvalue $\hat{z}_k$ of $\hat{S}$ used as an initial guess for a Newton iteration on a bordered system [Gov00, Ch. 3] to find the nearby eigenvalue z_k of T, the residuals $\sigma_{\min}T(z)$ for each initial guess and final iterate, and the absolute error between them. In every case the upper bound r_j on the error is comparatively generous but still tiny.

3.2.5 HIV I

This problem comes from modelling the spread of human immunodeficiency virus (HIV) on a cellular level [CR00, §3]. The linearized system [CR00, Eq. (3.2)] is of the form $\dot{v}(t) = A_0 v(t) + A_1 v(t-\tau)$, where $v(t)$ is a vector of length 3 representing concentrations of healthy cells, infected cells, and free HIV at time t, and the linearization is (as usual) computed at a known equilibrium. The corresponding matrix-valued function is

$$T(z) = zI - A_0 - A_1 e^{-\tau z}. \tag{3.22}$$

The equilibrium is asymptotically stable only if all eigenvalues of T are in the left half-plane.

$\hat{z}_k$	$\sigma_{\min} T(\hat{z}_k)$	$\sigma_{\min} T(z_k)$	$\lvert z_k - \hat{z}_k \rvert$	r_k
-4.484871e+00 + +8.958173e+01i	1.45e-11	1.21e-12	8.53e-14	1.00e-06
-3.210194e+00 + +7.493127e+01i	1.79e-12	9.22e-13	1.42e-14	1.00e-06
-2.999503e+00 + +7.222190e+01i	1.13e-11	5.08e-13	8.53e-14	1.00e-06
-2.766164e+00 + +6.909666e+01i	4.52e-13	5.47e-13	4.44e-16	1.00e-06
-2.459082e+00 + +6.475301e+01i	4.50e-12	8.66e-13	4.26e-14	1.00e-06
-2.209414e+00 + +6.099277e+01i	3.35e-13	3.35e-13	0.00e+00	1.00e-06
-2.030327e+00 + +5.814532e+01i	1.76e-12	4.00e-14	1.42e-14	1.00e-06
-1.971334e+00 + +5.717617e+01i	5.26e-12	1.89e-13	4.26e-14	1.00e-06
-1.834912e+00 + +5.486924e+01i	7.29e-12	6.32e-13	6.39e-14	1.00e-06
-1.704811e+00 + +5.257465e+01i	2.67e-12	2.81e-13	2.84e-14	1.00e-06
-1.394569e+00 + +4.664821e+01i	9.14e-14	5.50e-14	2.22e-16	1.00e-06
-1.291331e+00 + +4.450093e+01i	1.27e-12	4.94e-13	1.42e-14	1.00e-06
-1.226016e+00 + +4.308709e+01i	2.29e-12	3.77e-13	2.84e-14	1.00e-06
-1.005318e+00 + +3.792083e+01i	9.74e-13	1.01e-13	1.42e-14	1.00e-06
-9.068205e-01 + +3.537199e+01i	8.74e-13	8.13e-14	1.42e-14	1.00e-06
-7.461690e-01 + +3.076432e+01i	7.52e-13	9.47e-14	1.07e-14	1.00e-06
-6.160677e-01 + +2.645034e+01i	2.82e-13	1.04e-13	1.42e-14	1.00e-06
-5.639218e-01 + +2.450881e+01i	8.81e-14	8.81e-14	0.00e+00	1.00e-06
-4.098228e-01 + +1.755768e+01i	7.22e-13	1.74e-13	1.42e-14	1.00e-06
-3.541163e-01 + +1.423217e+01i	6.51e-14	1.44e-13	1.24e-14	1.00e-06
-3.431182e-01 + +1.347896e+01i	6.94e-13	2.38e-14	2.84e-14	1.00e-06
-2.618023e-01 + +5.229862e+00i	5.17e-13	1.43e-13	3.82e-14	1.00e-06
-2.656843e-01 + +5.892319e+00i	7.13e-14	5.82e-14	1.60e-14	1.00e-06
-2.781202e-01 + +7.636927e+00i	7.59e-14	6.08e-14	7.99e-15	1.00e-06

Table 3.2: Error table for the `hospital` problem.

Using the parameters in [CR00, Fig. 2(A4), Table 1], we have

$$A_0 = \begin{bmatrix} -4.3575 \times 10^{-2} & -5.2136 \times 10^{-3} & -6.2563 \times 10^{-3} \\ 0 & -2.6000 \times 10^{-1} & 0 \\ -4.2439 \times 10^{-2} & 1.2000 \times 10^{2} & -2.4063 \end{bmatrix} \tag{3.23}$$

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 3.5366 \times 10^{-2} & 0 & 5.2136 \times 10^{-3} \\ 0 & 0 & 0 \end{bmatrix}. \tag{3.24}$$

The matrix A_0 is diagonalizable and allows us to transform the problem to have tighter Gershgorin regions in the right half-plane (where the exponential term

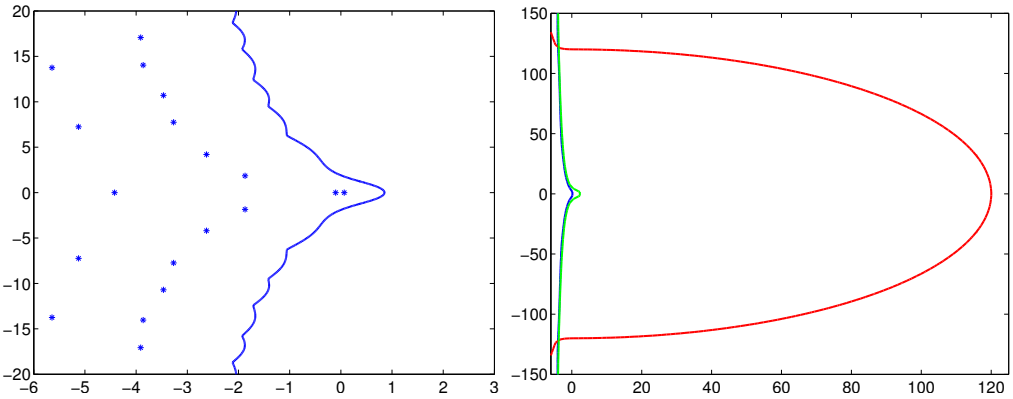

Figure 3.10: Localization region and envelope curves for HIV problem.

becomes small). If $A_0V = VD$, then

$$S(z) = zI - D - \tilde{A}_1 e^{-\tau z}, \qquad \tilde{A}_1 = V^{-1}A_1V. \tag{3.25}$$

Applying Theorem 2.1 to S gives the inclusion region shown in Figure 3.10 (left). The zeros of the diagonal entries of $S(z)$ are also plotted ($*$). Because they appear on the left of the boundary of the inclusion region, the inclusion region must be to the left of the plotted boundary and hence the eigenvalues of S (and T) are also to the left.

In Figure 3.10 we plot the envelope curves (see [MN07b, Prop. 1.10]) $|z| \leq \|A_0\|_2 + \|A_1\|_2 \exp(-\tau \Re e z)$ (red) and $|z| \leq \|D\|_2 + \|\tilde{A}_1\|_2 \exp(-\tau \Re e z)$ (green) for comparison. The former is far too generous for small imaginary parts but tighter than our blue curve for $|\Im m z| > 125$. The latter is comparable to our blue curve. Unlike other time delay problems we treat in this thesis, the matrix in the exponential term is not diagonalizable, and its eigenvalues are all zeros. It may be possible to find some scaling (such as the ones used in Section 3.5 for the delta potentials problem) or other transformation that will give tight inclusion regions in the left half-plane after an application of Theorem 2.1. If we are only concerned about stability, though, it is enough to compute the eigenvalues of T in the small region to the left of the blue curve and to the right of the imaginary

axis. We return to this problem in Section 3.6.

3.3 Using the Cholesky decomposition

Given a matrix-valued function consisting of a sum of a few terms, it is sometimes possible to diagonalize several simultaneously. In this section, we use the Cholesky decomposition [GL96] to do this for some examples where one term involves a Hermitian positive definite matrix.

3.3.1 Loaded string

The `loaded_string` problem from [BHM+13] is a rational eigenvalue problem of the form

$$T(z) = A - zB + \frac{z}{z-\sigma}C, \tag{3.26}$$

where $A, B, C \in \mathbb{R}^{20\times 20}$ and $\sigma = 1$. Both A and B are real, symmetric, tridiagonal, and positive, and C is zero except for $C_{20,20}$. The term with fastest growth in R is λB, and since B is symmetric positive definite we can apply a Cholesky decomposition to diagonalize it while having the freedom to further diagonalize another term. If $B = R^*R$ where R is the upper triangular Cholesky factor of B, then we can transform (3.26) into the matrix-valued function

$$\tilde{T}(z) = \tilde{A} - zI + \frac{z}{z-\sigma}\tilde{C}, \tag{3.27}$$

with the same eigenvalues as T, where $\tilde{A} = R^{-*}AR^{-1}$ and $\tilde{C} = R^{-*}CR^{-1}$. Then $\|\tilde{A}\|_2 \approx 5000$ whereas $\|\tilde{C}\|_2 \approx 70$, so it makes more sense to diagonalize $\tilde{A}$ than $\tilde{C}$

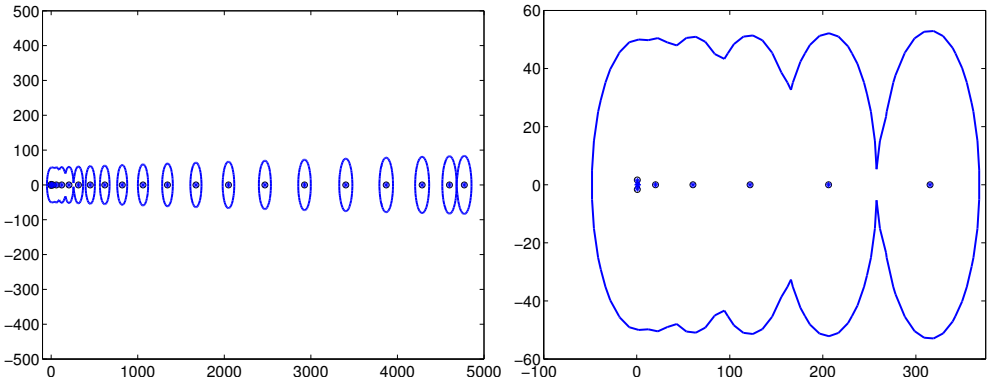

Figure 3.11: Localization regions and approximate eigenvalues for the `loaded_string` problem.

as the next step. If $\tilde{A}V = VD$, then

$$S(z) = D - zI + \frac{z}{z-\sigma}\hat{C}, \qquad \hat{C} = V^{-1}\tilde{C}V \tag{3.28}$$

has the same eigenvalues as (3.26). Applying Theorem 2.1 to S gives the localization regions and eigenvalue estimates ($*$) shown in Figure 3.11 (left). The actual eigenvalues of T are also plotted ($\circ$). Figure 3.11 (right) is a closer view of the left-most component. Due to the pole at $z = \sigma = 1$, this left-most component does not necessarily contain the same number of eigenvalues of S and $\hat{S}$, where $\hat{S}$ is the diagonal part of S.

Since A, B, and C are real, eigenvalues of T (and therefore of S) come in complex conjugate pairs. Furthermore, almost all connected components of the localization regions contain exactly one eigenvalue of $\hat{S}$. By the counting result in Theorem 2.1, S also has exactly one eigenvalue in each of those components, and therefore must be real. Hence, each localization region component containing one eigenvalue is actually a localization interval on the real line.

3.3.2 Butterfly III

We return to the `butterfly` problem from [BHM+13], of the form

$$T(z) = A_4 z^4 + A_3 z^3 + A_2 z^2 + A_1 z + A_0. \tag{3.29}$$

Out of the five matrices A_j, only A_0 and A_4 are symmetric positive definite. Therefore we can use the Cholesky decomposition on either. Since A_0 and A_4 have roughly the same norm, it makes more sense to prioritize making the term with fastest growth diagonal. Letting $A_4 = R^* R$, where R is the upper triangular Cholesky factor, we can transform (3.29) into another matrix-valued function with the same eigenvalues, namely

$$\tilde{T}(z) = I z^4 + B_3 z^3 + B_2 z^2 + B_1 z + B_0, \qquad B_j = R^{-*} A_j R^{-1}. \tag{3.30}$$

Now we are free to diagonalize any of the other terms. Since all the B_j's have roughly the same norm, it would seem likely that diagonalizing the term with largest growth, namely the z^3 term, would give the best inclusion regions out of the four choices. Trying each confirms this supposition. If $B_3 V = VD$, then conjugating (3.30) by V gives the matrix-valued function

$$S(z) = I z^4 + D z^3 + C_2 z^2 + C_1 z + C_0, \qquad C_j = V^{-1} B_j V \tag{3.31}$$

with the same eigenvalues as (3.29). Splitting (3.31) as $S(z) = \hat{S}(z) + E(z)$, into the sum of its diagonal and off-diagonal parts, respectively, and applying Theorem 2.1 gives the localization region (red, dashed) and eigenvalues of $\hat{S}$ $(*)$ plotted in red in Figure 3.12 (left). Plotted in blue are the naïve localization regions computed in Section 3.1 as well as the eigenvalues of the diagonal part of (3.29). The black circles are the true eigenvalues of (3.29). In Figure 3.12 (right) we have added the localization regions and eigenvalue estimates obtained by diag-

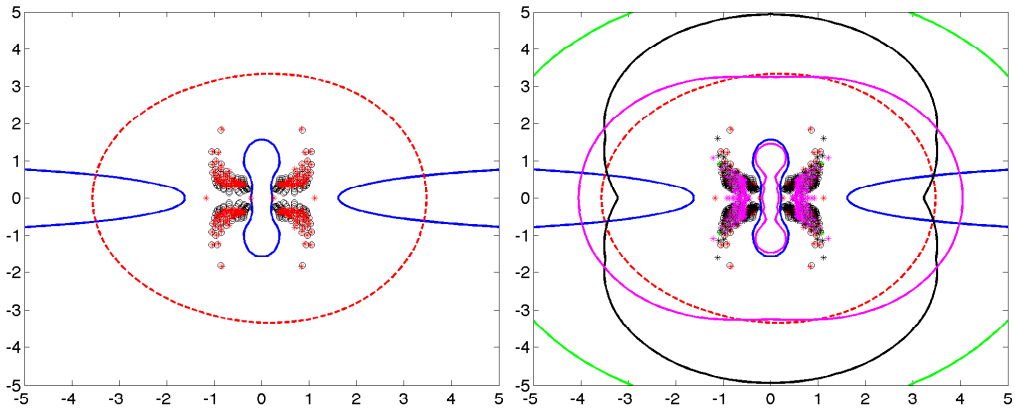

Figure 3.12: Localization regions for `butterfly` problem obtained with Cholesky decomposition of A_4 and diagonalization of another term.

onalizing B_2 (green), B_1 (black), or B_0 (magenta) instead of B_3 for comparison.

3.4 Dropping terms

So far we have performed similarity transformations on a matrix-valued function $T(z)$ to obtain another matrix-valued function $\tilde{T}(z)$ with the same eigenvalues, and split $\tilde{T}$ into its diagonal and off-diagonal parts $D(z)$ and $E(z)$, respectively. Occasionally a splitting where the diagonal of $E(z)$ is nonzero may be more practical.

3.4.1 Hadeler

The `hadeler` problem from [BHM+13] is of the form

$$T(z) = (e^z - 1)B + z^2A - \alpha I \tag{3.32}$$

where $A, B \in \mathbb{R}^{8\times 8}$ are symmetric and positive definite and $\alpha = 100$. This example has been analyzed in [BH13] using both the Gershgorin generalization Theorem 2.1 and the pseudospectral localization result in Theorem 2.5 with excellent results. Here we focus on one aspect of the analysis that demonstrates a part of Theorem 2.1 that we have not had occasion to use in the other examples in this thesis.

Suppose we take advantage of our ability to diagonalize both A and B simultaneously by performing a Cholesky decomposition $A = R^T R$ and subsequently diagonalizing $R^{-T} B R^{-1}$. This leads to the matrix-valued function

$$S(z) = (e^z - 1)D + z^2 I - F, \qquad \left(R^{-T} B R^{-1}\right) V = VD, \quad F = V^{-1}\left(R^{-T} R^{-1}\right) V. \tag{3.33}$$

Since the terms which become large away from the origin (namely the e^z and the z^2 terms) are diagonal, we know that the inclusion regions that come from splitting $S(z)$ into diagonal $(e^z - 1)D + z^2 I - D_F$ and off-diagonal $F - D_F$ (where D_F is the diagonal part of F) and applying Theorem 2.1 will be tight far from the origin. Indeed, this is borne out by looking at the inclusion region (thick red line) in Figure 3.13 (left) where a pseudospectral plot for T also indicates the eigenvalues of T. In fact, the result is better than expected, because even the components of the inclusion region near the origin are very tight. Most of the components off the real line are so tiny they are not visible in the figure, with only the components near $6.5 \pm 5i$ showing.

There is a drawback to using this diagonal/off-diagonal splitting in this case, which is that the eigenvalues of the diagonal part of S, namely $(e^z - 1)D + z^2 I - D_F$, are hard to compute analytically, thus making it difficult to use the counting result in Theorem 2.1 or to use the eigenvalues of the diagonal part as approximations to the eigenvalues of T. Most importantly, suppose we would like all

the eigenvalues of T in the region pictured in Figure 3.13 (left or right) and want to use inclusion region and pseudospectral plots as a visual guide. As we have already pointed out, it is quite possible that with a given choice of mesh of this region, a component of the Gershgorin region surrounding an eigenvalue may not appear. Also, an eigenvalue may not be contained in a component of any of the plotted ε-pseudospectra, or a component of an ε-pseudospectrum could surround a group and we could miss one or more of said group by not knowing how many we should expect there. The only way we can be sure to avoid missing eigenvalues of T is by also having the eigenvalues of the diagonal summand in whatever splitting we use in Theorem 2.1.

Recall that Theorem 2.1 allows arbitrary splittings of e.g. S as long as the first summand is a diagonal matrix. For instance, we may choose $S(z) = \hat{S}(z) + E(z)$ where

$$\hat{S}(z) = e^z D + z^2 I, \qquad E(z) = -D - F, \tag{3.34}$$

the latter being a constant matrix-valued function with nonzero diagonal entries. The inclusion regions we get with this splitting (thick blue line in Figure 3.13 (right)) are much more generous than the ones derived using the first splitting (Figure 3.13 (left)). However, we can write the eigenvalues of $\hat{S}$ in closed form using the Lambert W function [CGH+96]. In particular, $e^z D_{jj} + z^2 = 0$ is equivalent to $z^2 = -D_{jj} e^z$, which is equivalent to $z = \pm i \sqrt{D_{jj}} e^{z/2}$. Rewriting this as $(-z/2)e^{-z/2} = \pm i \sqrt{D_{jj}}/2$ exhibits the original equation in the form $we^w = x$ for $x = \pm i \sqrt{D_{jj}}/2$, whose solutions w can be computed using e.g. `lambertw` in MATLAB. The eigenvalues of $\hat{S}$ computed this way are plotted in Figure 3.13 (right) with blue stars. Clearly they do not approximate the real eigenvalues of T very well, but are better and better approximations as we move away from the real axis. In fact, it is shown in [BH13] that if $\hat{\lambda}$ is an eigenvalue of $\hat{S}$ with

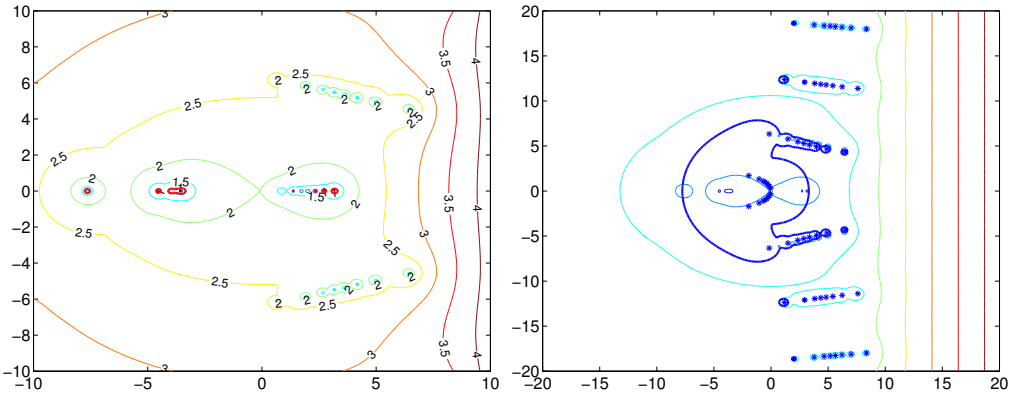

Figure 3.13: Inclusion regions for the `hadeler` problem using different splittings.

$|\hat{\lambda}| > 16.3$, then the error in $\hat{\lambda}$ approximating a true eigenvalue of T is $O(|\hat{\lambda}|^{-2})$.

As for the union U of inclusion region components in $|\Im\mathrm{m} z| < 10$, there are many ways of obtaining the eigenvalues of T within it. An approach taken in [BH13] is to approximate T by a Chebyshev polynomial such that the approximation is excellent near the real axis, and reason about pseudospectral inclusion regions using Theorem 2.5. On the other hand, knowing the number of eigenvalues of $\hat{S}$ in U gives us the number of eigenvalues of T in U and U is well-separated from the rest of the inclusion regions, so we could simply use an algorithm like Beyn's [Bey12]. In particular, $\hat{S}$ has—and therefore T also has—32 eigenvalues in U, which is greater than the number of rows in T, so we must use Beyn's Integral Algorithm 2 [Bey12, p. 3860]. As inputs, we choose the contour $|z| = 10$ shown in Figure 3.14 (left), $N = 500$, $\ell = 8$ (as directed in Step 1 of the algorithm), $K = 4$ (so that $K\ell \geq 32$), `tolrank` $= 10^{-8}$, `tolres` $= 10^{-6}$, and `maxcond` $= 10$. The eigenvalues λ computed this way are plotted in Figure 3.14 (right) with black dots and tabulated along with their residuals $\sigma_{\min} T(\lambda)$ in Table 3.3.

	λ	$\sigma_{\min} T(\lambda)$
1	+2.335425e+00 -1.074084e-13i	3.93e-11
2	+2.731077e+00 -1.284750e-13i	3.75e-10
3	+3.182596e+00 -1.776527e-13i	5.91e-10
4	-3.571756e+00 +1.184595e-14i	1.33e-12
5	-3.627468e+00 +1.592141e-14i	2.33e-12
6	-3.702762e+00 +5.605691e-15i	8.74e-13
7	-3.491853e+00 +3.655253e-14i	1.23e-11
8	-3.801275e+00 +9.039860e-15i	1.92e-12
9	-3.968169e+00 +7.906693e-15i	2.82e-12
10	-4.521556e+00 +4.282154e-15i	3.09e-13
11	+3.178272e+00 +5.492525e+00i	1.33e-11
12	+3.621948e+00 +5.359316e+00i	2.29e-11
13	+2.688852e+00 +5.638766e+00i	1.44e-11
14	+4.187385e+00 +5.191003e+00i	2.31e-11
15	+1.928091e+00 +5.867287e+00i	1.10e-11
16	+5.008011e+00 +4.952609e+00i	2.71e-11
17	+7.222701e-01 +6.190483e+00i	7.39e-12
18	+6.460619e+00 +4.569605e+00i	2.23e-10
19	+3.178272e+00 -5.492525e+00i	1.06e-11
20	+3.621948e+00 -5.359316e+00i	1.08e-11
21	+2.688852e+00 -5.638766e+00i	1.59e-11
22	+4.187385e+00 -5.191003e+00i	2.76e-11
23	+1.928091e+00 -5.867287e+00i	1.15e-11
24	+5.008011e+00 -4.952609e+00i	2.49e-11
25	+7.222701e-01 -6.190483e+00i	3.51e-12
26	+6.460619e+00 -4.569605e+00i	2.27e-10
27	-7.642558e+00 -4.440892e-15i	3.61e-12
28	+2.007944e+00 +8.441774e-14i	2.18e-11
29	+1.726304e+00 +4.224966e-14i	5.09e-12
30	+1.394724e+00 +1.066024e-13i	6.19e-10
31	+8.849615e-01 +1.537867e-13i	2.75e-09
32	+2.174614e-01 -1.897609e-14i	2.71e-08

Table 3.3: Error table for the `hadeler` problem.

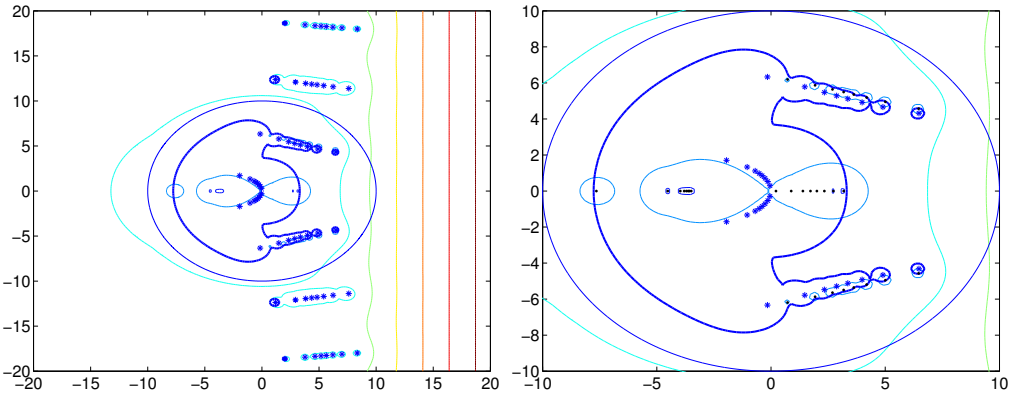

Figure 3.14: Using Beyn's Integral Algorithm 2 to compute eigenvalues for the `hadeler` problem.

3.5 Using block structure

From time to time a matrix-valued function does not yield to our diagonalization tricks, but can be manipulated in such a way that its *block* diagonal is dominant. Then, it is worth trying the slightly more difficult to use block version of Theorem 2.1, namely Theorem 2.2. The first example in this section is of this type. The second example demonstrates that even if Theorem 2.1 is successfully applied to a problem, Theorem 2.2 may be even more effective.

3.5.1 Delta potentials

We might see a matrix like

$$\begin{bmatrix} 1 & -1 & -1 & 0 \\ 1/2 - ik & 1/2 - ik & 1/2 + ik & 0 \\ 0 & e^{ik} & e^{-ik} & -e^{ik} \\ 0 & (1/2 + ik)e^{ik} & (1/2 - ik)e^{-ik} & (1/2 - ik)e^{ik} \end{bmatrix} \tag{3.35}$$

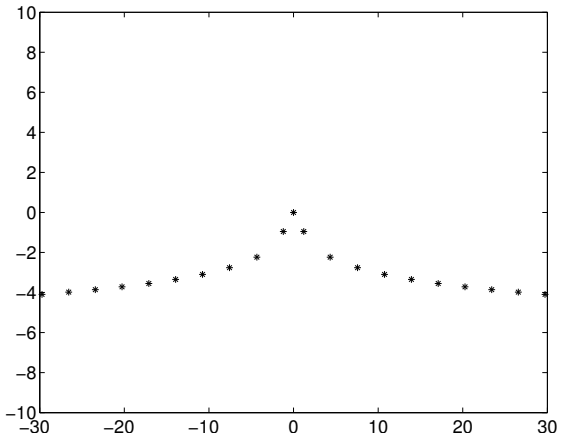

Figure 3.15: Square roots of resonances for $V(x) = \delta(x) + \delta(x-1)$.

in a physics student's homework. It comes from computing solutions ψ for the one-dimensional Schrödinger equation $(-\Delta + V)\psi = k^2\psi$, where the potential function is $V(x) = \delta(x) + \delta(x-1)$. See Appendix A for a derivation.

If the determinant of the matrix (3.35) is zero, then k^2 is called a resonance for $V(x) = \delta(x) + \delta(x-1)$. By computing the determinant of this simple 4x4 matrix we find that the square roots of resonances satisfy $k = -i(1/2 - w)$ where $we^w = \pm e^{1/2}/2$. The latter equation can be solved using the Lambert W function [CGH+96], and the (infinitely many) solutions k with $-30 < \Re\mathfrak{e}k < 30$ are plotted in Figure 3.15.

If we switch to the case where the delta functions have different strengths and the second delta function is at $p > 0$, i.e., $V(x) = L_1\delta(x) + L_2\delta(x-p)$ with $L_1 \neq L_2$, then the matrix defining the resonances is

$$\begin{bmatrix} 1 & -1 & -1 & 0 \\ L_1/2 - ik & L_1/2 - ik & L_1/2 + ik & 0 \\ 0 & e^{ikp} & e^{-ikp} & -e^{ikp} \\ 0 & (L_2/2 + ik)e^{ikp} & (L_2/2 - ik)e^{-ikp} & (L_2/2 - ik)e^{ikp} \end{bmatrix}, \tag{3.36}$$

and setting the determinant of (3.36) equal to zero shows that the resonances for

$V(x) = L_1\delta(x) + L_2\delta(x-p)$ satisfy $(L_1 - 2ik)(L_2 - 2ik) = L_1L_2e^{i2kp}$. Now that $L_1 \neq L_2$, it's not obvious how to massage this into the form $w(k)e^{w(k)} = x$ (or any form) that lends itself to closed form solutions. The analysis of the $L_1 = L_2 = p = 1$ case suggests there are infinitely many solutions of $(L_1 - 2ik)(L_2 - 2ik) = L_1L_2e^{i2kp}$ and we might be able to find them on a curve like the one in Figure 3.15. But is this always true, even if $|L_1 - L_2|$ is large or the signs differ? Further, what happens if the potential $V(x)$ is a sum of n delta functions, leading to a $2n \times 2n$ matrix? We will turn to the theorems from Chapter 2 for a partial answer.

Let us first consider the case of a two-delta potential $V(x) = L_1\delta(x) + L_2\delta(x-p)$ with $p > 0$ and L_1 not necessarily equal to L_2. The corresponding matrix-valued function whose eigenvalues we want is

$$T(k) = \begin{bmatrix} 1 & -1 & -1 & 0 \\ \frac{L_1}{2} - ik & \frac{L_1}{2} - ik & \frac{L_1}{2} + ik & 0 \\ 0 & e^{ikp} & e^{-ikp} & -e^{ikp} \\ 0 & \left(\frac{L_2}{2} + ik\right)e^{ikp} & \left(\frac{L_2}{2} - ik\right)e^{-ikp} & \left(\frac{L_2}{2} - ik\right)e^{ikp} \end{bmatrix}. \tag{3.37}$$

If we split this into a sum $\sum_j p_j(k)A_j$, each of the matrices A_j is low rank. Therefore, diagonalizing any one of the terms is of limited usefulness in applying Theorem 2.1. However, by appropriate scalings, we can transform $T(k)$ into a matrix-valued function with the same eigenvalues and useful block structure. Our guiding thought is that $e^{ikp} = e^{ipx}e^{-yp}$ ($k = x + iy$) is large in the lower half-plane and small in the upper half-plane.

First, scale $T(k)$ on the left by $\mathrm{diag}([L_1/2 - ik, 1, L_2/2 - ik, 1])$ and on the right

by $\mathrm{diag}([1, 1, \exp(ikp), \exp(-ikp)])$ to obtain

$$S(k) = \begin{bmatrix} \frac{L_1}{2} - ik & -\left(\frac{L_1}{2} - ik\right) & -\left(\frac{L_1}{2} - ik\right)e^{ikp} & 0 \\ \frac{L_1}{2} - ik & \frac{L_1}{2} - ik & \left(\frac{L_1}{2} + ik\right)e^{ikp} & 0 \\ 0 & \left(\frac{L_2}{2} - ik\right)e^{ikp} & \frac{L_2}{2} - ik & -\left(\frac{L_2}{2} - ik\right) \\ 0 & \left(\frac{L_2}{2} + ik\right)e^{ikp} & \frac{L_2}{2} - ik & \frac{L_2}{2} - ik \end{bmatrix}. \tag{3.38}$$

The left diagonal scaling introduced the eigenvalues $-iL_1/2$ and $-iL_2/2$ for S which T may not have had. Aside from that exception, S has the same spectrum as T. If we partition $S(z)$ into 2×2 blocks, the off-diagonal blocks will be small compared to the diagonal blocks wherever $\exp(ikp)$ is small, which it is in the upper half-plane. Therefore, in the upper half-plane, the off-diagonal blocks are small in norm compared to the diagonal blocks. This gives us the opportunity to obtain tight inclusion regions in the upper half-plane by using Theorem 2.2. To simplify notation, let $X = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ so that the diagonal blocks of S are $(L_1/2 - ik)X$ and $(L_2/2 - ik)X$. Note that $\|X^{-1}\|_\infty^{-1} = 1$. Then Theorem 2.2, using the infinity norm, shows that eigenvalues of S are confined to the union $G_1 \cup G_2$ where

$$G_1 = \left\{ k : |L_1/2 - ik| \le |e^{ikp}| \max\left(|L_1/2 - ik|, |L_1/2 + ik|\right) \right\} \tag{3.39}$$

$$= \left\{ k = x + iy : e^{yp} \le \max\left(1, \frac{|L_1/2 + ik|}{|L_1/2 - ik|}\right) \right\}, \tag{3.40}$$

$$G_2 = \left\{ k = x + iy : e^{yp} \le \max\left(1, \frac{|L_2/2 + ik|}{|L_2/2 - ik|}\right) \right\}. \tag{3.41}$$

Clearly both G_1 and G_2 include a subset of $\mathbb{C}$ with sufficiently negative real part and cannot include complex numbers with large positive imaginary part. Therefore the boundary of $G_1 \cup G_2$ bounds the spectrum of S (and T) from above. This boundary appears as the thick black line in each plot in Figures 3.16 and 3.17.

For the second transformation, we can swap the second and third columns

of $T(k)$ and then multiply on the right by $\mathrm{diag}([\exp(ikp), \exp(ikp), 1, 1])$, to obtain

$$R(k) = \begin{bmatrix} e^{ikp} & -e^{ikp} & -1 & 0 \\ \left(\frac{L_1}{2} - ik\right) e^{ikp} & \left(\frac{L_1}{2} + ik\right) e^{ikp} & \left(\frac{L_1}{2} - ik\right) & 0 \\ 0 & 1 & e^{ikp} & -e^{ikp} \\ 0 & \left(\frac{L_2}{2} - ik\right) & \left(\frac{L_2}{2} + ik\right) e^{ikp} & \left(\frac{L_2}{2} - ik\right) e^{ikp} \end{bmatrix}. \tag{3.42}$$

Clearly, if $T(k)v = 0$, then $R(k)\tilde{v} = 0$ for $\tilde{v}$ a permuted and scaled version of v, and vice versa. Therefore R and T have the same eigenvalues. Partitioning R into 2×2 blocks, it is clear that the diagonal blocks are multiples of e^{ikp} and the off-diagonal blocks do not involve e^{ikp}. Since the diagonal blocks are large in norm compared to the off-diagonal blocks in the lower half-plane, where e^{ikp} is large, Theorem 2.2 can give tight inclusion regions for the spectrum of R in the lower half-plane. Similar to the analysis we did before, the boundary of the block Gershgorin region for R derived from Theorem 2.2 bounds the spectrum of R (and T) from below. The bound from below is shown as the thick blue line in each plot in Figures 3.16 and 3.17.

Pseudospectral plots suggesting the locations of eigenvalues of (3.37) corresponding to a potential $\delta(x) + \delta(x - 1)$ (top left), $-\delta(x) - \delta(x - 1)$ (top right), $20\delta(x) + 20\delta(x - 1)$ (bottom left), and $100\delta(x) + 100\delta(x - 1)$ (bottom right) are shown in Figure 3.16 along with the upper bounds (thick black line) and lower bounds (thick blue line) we derived by transforming (3.37) and applying Theorem 2.2. Since we can compute the eigenvalues in this case, they are plotted as well ($*$).

The transformations we have done in this section can be generalized to an arbitrary number of delta potentials with arbitrary real strengths. For instance, if $V(x) = L_1\delta(x) + L_2\delta(x - p_2) + L_3\delta(x - p_3)$ is a potential with three resonances at

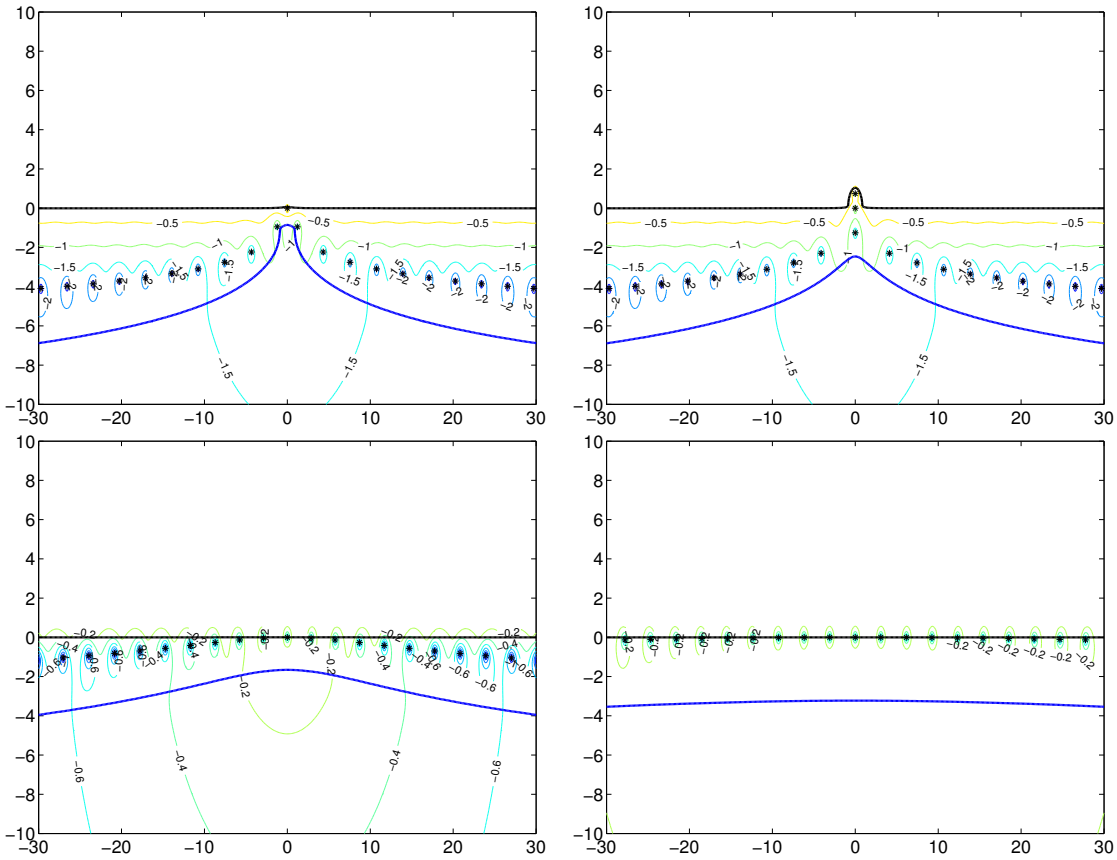

Figure 3.16: Inclusion regions and pseudospectra for some problems with 2 delta potentials of equal strength.

$0 < p_2 < p_3$, then its resonances are the eigenvalues of

$$T_3(k) = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ \frac{L_1}{2} - ik & \frac{L_1}{2} - ik & \frac{L_1}{2} + ik & 0 & 0 & 0 \\ 0 & e^{ikp_2} & e^{-ikp_2} & -e^{ikp_2} & -e^{-ikp_2} & 0 \\ 0 & \left(\frac{L_2}{2} + ik\right) e^{ikp_2} & \left(\frac{L_2}{2} - ik\right) e^{-ikp_2} & \left(\frac{L_2}{2} - ik\right) e^{ikp_2} & \left(\frac{L_2}{2} + ik\right) e^{-ikp_2} & 0 \\ 0 & 0 & 0 & e^{ikp_3} & e^{-ikp_3} & -e^{ikp_3} \\ 0 & 0 & 0 & \left(\frac{L_3}{2} + ik\right) e^{ikp_3} & \left(\frac{L_3}{2} - ik\right) e^{-ikp_3} & \left(\frac{L_3}{2} - ik\right) e^{ikp_3} \end{bmatrix}. \tag{3.43}$$

If we define S_3 by multiplying T_3 by $\mathrm{diag}([L_1/2 - ik, 1, L_2/2 - ik, 1, L_3/2 - ik, 1])$ on the left and $\mathrm{diag}([1, 1, \exp(ikp_2), \exp(-ikp_2), \exp(ikp_3), \exp(-ikp_3])$, then S_3 will have block structure similar to S. If we define R_3 by swapping

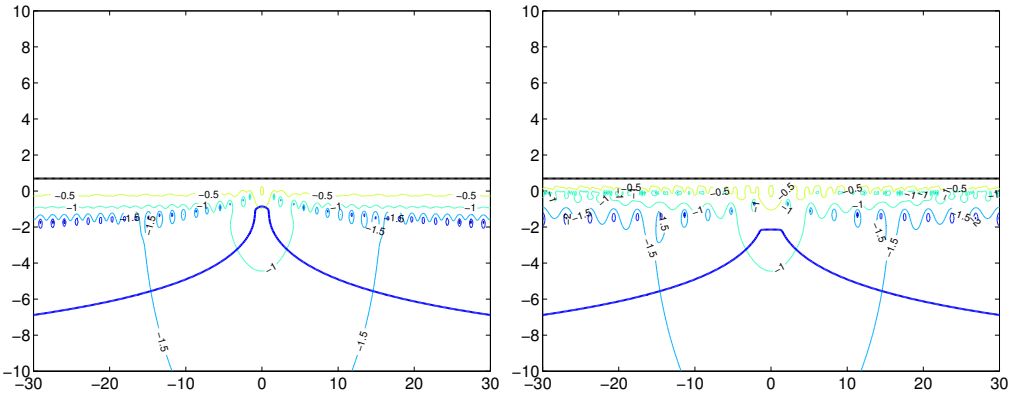

Figure 3.17: Inclusion regions and pseudospectra for problems with 3 and 10 delta potentials of differing strengths.

columns 2 and 3, then swapping columns 4 and 5, and finally applying $\mathrm{diag}([\exp(ikp_2), \exp(ikp_2), \exp(-ikp_1), \exp(ikp_3), \exp(-ikp_2), 1])$ on the right, then R_3 will have block structure similar to R. Applying Theorem 2.2 to S_3 and R_3 give bounds on the spectrum of T from above and below, respectively.

Pseudospectra for matrix-valued functions associated to $V(x) = \delta(x) + 2\delta(x - 1/2) + 3\delta(x-1)$ and $V(x) = \sum_{n=0}^{4}(\delta(x-2n)+20\delta(x-2n-1))$ are pictured in Figure 3.17 (left) and (right), respectively. The blue curves pictured are the boundaries of the block Gershgorin regions for the associated matrix-valued functions "R" and thus gives bounds from below on the spectrum of the associated "T". Similarly, the black curves are derived from the corresponding matrix-valued functions analogous to "S" and give bounds from above. These inclusion regions are not tight, but at least give an idea of where in the complex plane we should focus our attention.

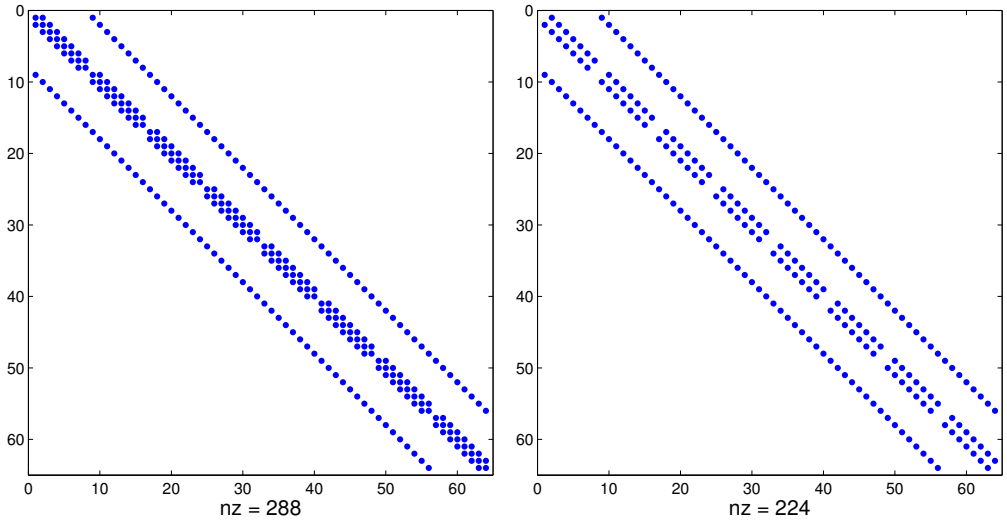

Figure 3.18: Structure of matrices in `butterfly` problem.

3.5.2 Butterfly IV

We return for a final time to the `butterfly` problem from [BHM+13] in order to present the best localization region we have found. Recall that this problem is to find the eigenvalues of matrix-valued function $T(z) = A_4z^4 + A_3z^3 + A_2z^2 + A_1z + A_0$. We have mentioned in Section 3.2 that each A_j is a sum of Kronecker products. As a result, the structure of each A_j is block tridiagonal after partitioning into 8×8 blocks. See Figure 3.18 for the structure of A_0, A_2, and A_4 (left) and the structure of A_1 and A_3 (right). The off-diagonal blocks are themselves multiples of the identity, and are smaller in norm than the diagonal blocks. This suggests we use Theorem 2.2 with $T(z) = \hat{T}(z) + E(z)$ the splitting into the 8×8 block diagonal part of T and the rest, respectively. Indeed, Figure 3.19 (left) shows the resulting localization regions, and they are much tighter than the localization regions we have computed for the `butterfly` problem in previous sections (compare with Figure 3.19 (right), copied from Figure 3.12 (right) and described in Section 3.2). The actual eigenvalues of T ($\circ$) and eigenvalues of $\hat{T}$ ($*$) are also

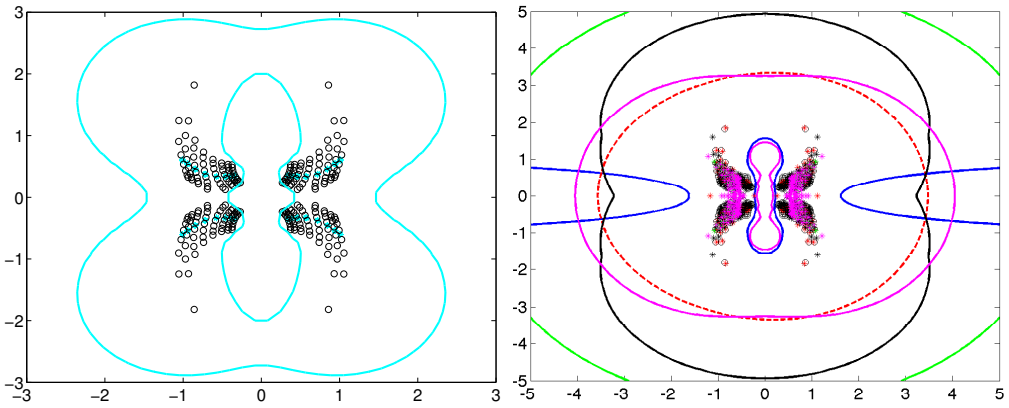

Figure 3.19: Localization region for `butterfly` problem obtained by applying the nonlinear block Gershgorin theorem.

plotted.

The drawback of Theorem 2.2 in comparison to Theorem 2.1 is that the eigenvalues of a block diagonal matrix-valued function are harder to compute than eigenvalues of a diagonal matrix-valued function, which could make it more difficult to use the counting aspect of the theorems. However, this example shows that using Theorem 2.2 may well be worth the extra effort.

3.6 Using approximations

When Theorems 2.1 and 2.2 fail to help localize the eigenvalues of a genuinely nonlinear matrix-valued function, it is time to construct an approximation and use Theorem 2.5.

3.6.1 Gun

Here we analyze the `gun` problem from [BHM+13]. It is a matrix-valued function of the form

$$F(\lambda) = K - \lambda M + i\sqrt{\lambda}W_1 + i\sqrt{\lambda - \sigma_2^2}W_2, \tag{3.44}$$

where $K, M, W_1, W_2 \in \mathbb{R}^{9956\times 9956}$ are sparse and symmetric, $K \geq 0$, $M > 0$, $\sigma_2 = 108.8774$, and the principal branch of the square root is used.

The first thing that should be mentioned is, contrary to some statements made in the literature (e.g. "there is no efficient transformation to convert [it] to the polynomial eigenvalue problem" from [LBLQL10]), there is a very useful change of variable that makes (3.44) a polynomial eigenvalue problem (see Appendix B). We will use this change of variable to obtain exact eigenvalues of (3.44) for comparison, but proceed with the localization process as if we do not have exact eigenvalues available.

In [Lia07], the principal square roots of some eigenvalues of (3.44) are tabulated. We will also focus on the square roots $z = \sqrt{\lambda}$, where the principal branch is used. Changing to the new variable z, we define

$$T(z) = F(z^2) = K - z^2M + izW_1 + iz\sqrt{1 - \left(\frac{\sigma_2}{z}\right)^2}W_2. \tag{3.45}$$

Now, according to [Lia07, p. 59], we are only interested in eigenvalues of T which are near but to the right of 146.71. Under these conditions, $|\sigma_2/z| < 1$, so the Taylor expansion

$$\sqrt{1 - \left(\frac{\sigma_2}{z}\right)^2} = 1 - \frac{1}{2}\frac{\sigma_2^2}{z^2} - \frac{1}{8}\frac{\sigma_2^4}{z^4} - \frac{1}{16}\frac{\sigma_2^6}{z^6} - \ldots \tag{3.46}$$

converges reasonably quickly. Truncating it provides us with a rational approximation, and a matrix-valued function $\hat{T}$ that approximates T in the region of

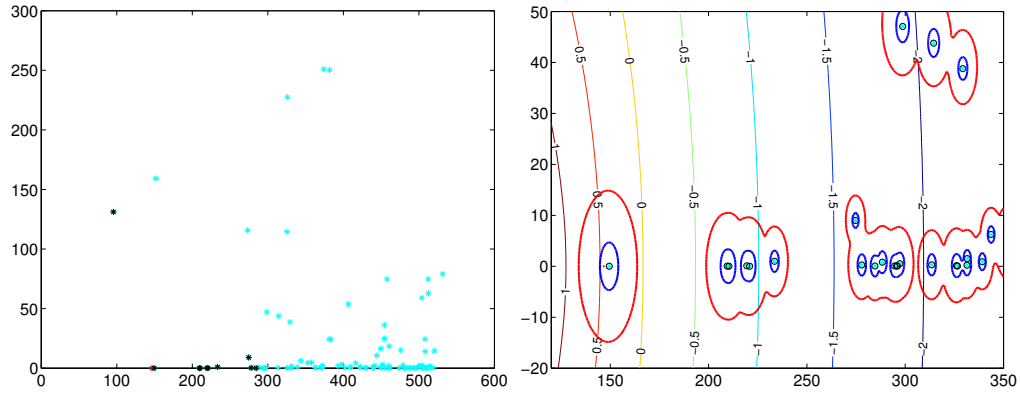

Figure 3.20: Exact and approximate eigenvalues and localization regions for `gun` problem.

interest. We define

$$\hat{T}(z) = K - z^2 M + izW_1 + izW_2 - \frac{i}{2}\frac{\sigma_2^2}{z}W_2 - \frac{i}{8}\frac{\sigma_2^4}{z^3}W_2 - \frac{i}{16}\frac{\sigma_2^6}{z^5}W_2, \tag{3.47}$$

so that $T(z) = \hat{T}(z) + E(z)$, with $\|E(z)\|_2$ reasonably small in the region of interest because $\|W_2\|_2 \approx 3.2$ is not too large. The eigenvalues of T near 146.71, computed via a companion linearization of the matrix polynomial derived in Appendix B, are plotted in Figure 3.20 (left), with the 10 eigenvalues of T closest to 146.71 plotted in black for easy comparison with [Lia07, Tables 6.3, 6.5]. The shift 146.71 is plotted in red. Figure 3.20 (right) contains eigenvalues of T ($*$), eigenvalues of $\hat{T}$ ($\circ$), contours of $\log_{10}\|E(z)\|_2$, and the 2-norm ε-pseudospectrum of T for $\varepsilon = 10^1$ (red) and $\varepsilon = 10^{0.5}$ (blue). Since we are using the 2-norm, the pseudospectra of T are the contours of $\sigma_{\min}(T(z))$, but we have actually plotted $\log_{10}(\sigma_{\min}(T(z)))$ contours.

The three conclusions we can draw from Figure 3.20 (right) by using Theorem 2.5 are as follows. First, since the pictured components of the 10^1-pseudospectrum of T are contained in the region where $\|E(z)\|_2 < 10^1$, each of those components contains exactly the same number of eigenvalues of T and eigenvalues of $\hat{T}$. Second, all but the leftmost pictured component of the $10^{0.5}$-

pseudospectrum of T is in the region where $\|E(z)\|_2 < 10^{0.5}$. Therefore all pictured components of the $10^{0.5}$-pseudospectrum (except the leftmost one) contain the same number of eigenvalues of T and $\hat{T}$. Third, every component of the 2-norm 10^1-pseudospectrum in the region Ω_{10} where $\|E(z)\|_2 < 10^1$ must necessarily also contain an eigenvalue of $\hat{T}$ (see Proposition 2.2 and the counting result in Theorem 2.5). Thus, since all eigenvalues of $\hat{T}$ in the pictured rectangle have been plotted, we can be assured by examining Figure 3.20 (right) that we have not missed some tiny component of the 2-norm 10^1-pseudospectrum of T which was too small to be detected with the mesh we used. Since every eigenvalue of T is contained in every ε-pseudospectrum, and there are no other components of the 10^1-pseudospectrum in the pictured region, there are no eigenvalues of T in the complement of the 10^1-pseudospectrum in the pictured part of Ω_{10}. Therefore we have localized and counted all eigenvalues of T within the intersection of $[120, 350] \times [-20, 50]$ with Ω_{10} in the complex plane.

3.6.2 HIV II

We return to the HIV problem from Section 3.2.5, of the form

$$T(z) = zI - A_0 - A_1 e^{-\tau z}, \tag{3.48}$$

where A_0 is diagonalizable but A_1 is not. Because all the eigenvalues of A_1 are zero and A_1 is not diagonalizable, the best localization region we were able to find in Section 3.2.5 was to the left of the blue curve in Figure 3.21 (top left), where zeros of the diagonal entries of (3.25) are plotted as well ($*$).

As discussed in Section 3.2.5, to study stability we need only compute eigenvalues of T to the left of the blue curve and to the right of the imaginary axis.

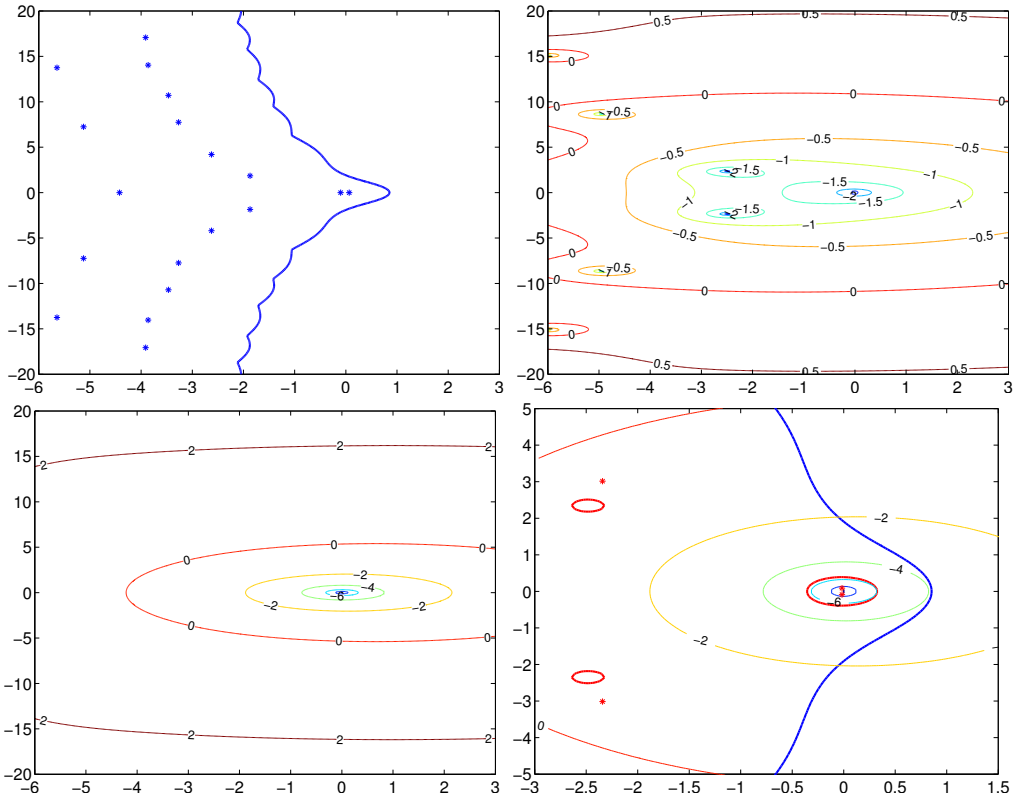

Figure 3.21: Deriving pseudospectral inclusion for the HIV problem.

We will use Theorem 2.5 to do this. First, we look at the 2-norm pseudospectra for T as defined in Section 2.3. These are pictured in Figure 3.21 (top right) as labeled contours of $\log_{10} \sigma_{\min} T(z)$.

Now, to use Theorem 2.5, we need to find a matrix-valued function $\hat{T}$ such that the eigenvalues of $\hat{T}$ are easy to compute and such that $\|T - \hat{T}\|_2$ is small in the region of interest (to the left of the blue curve but to the right of the imaginary axis) and perhaps a little to the left as well, since the pseudospectral plot for T shows there is an eigenvalue for T near the origin. There are many ways we can choose $\hat{T}$, but in this case we have found that a Taylor polynomial works quite well and we do not need to resort to anything more complicated. Specifi-

cally, taking $\exp(-\tau z) \approx \sum_{j=0}^{4}(-\tau z)^j/j!$, we define

$$\hat{T}(z) = zI - A_0 - A_1\left(1 - \tau z + \frac{\tau^2}{2!}z^2 - \frac{\tau^3}{3!}z^3 + \frac{\tau^4}{4!}z^4\right) \tag{3.49}$$

which becomes a matrix polynomial after grouping by powers of z. A plot of $\log_{10}\|T(z) - \hat{T}(z)\|_2$ is shown in Figure 3.21 (bottom left).

The 10^{-2}-pseudospectrum of T is plotted in Figure 3.21 (bottom right) as red contours. Clearly the right-most component of the 10^{-2}-pseudospectrum of T is in the region where $\|T - \hat{T}\|_2 < 10^{-2}$. Therefore, by Theorem 2.5, this right-most component contains the same number of eigenvalues of $\hat{T}$ and T, and furthermore T has no other eigenvalues in the region where $\|T - \hat{T}\|_2 < 10^{-2}$. Since the latter encloses the region of interest (left of the blue curve, right of the imaginary axis), and since $\hat{T}$ has two eigenvalues (red stars) in the right-most component of the 10^{-2}-pseudospectrum of T, we need only compute two eigenvalues of T to answer the question of whether the eigenvalues of T are all in the left half-plane.

The two right-most eigenvalues of $\hat{T}$ are $-0.0167 \pm 0.0886i$. We use these as initial guesses for Kressner's algorithm [Kre09] applied to T with parameter $\ell = 1$, and compute that the two nearby eigenvalues $\lambda_{1,2}$ of T differ from the initial guesses by 0.48×10^{-8} (and satisfy $\sigma_{\min}T(\lambda_{1,2}) \approx 2 \times 10^{-17}$). Hence, these eigenvalues of T are in the left half-plane, and furthermore all eigenvalues of T are in the left half-plane. This gives a positive answer to the asymptotic stability question discussed in Section 3.2.5.

One final comment refers to Figure 3.21 (bottom right). The reader will notice that the components of the 10^{-2}-pseudospectrum of T (red contours) on the left contain no eigenvalues of $\hat{T}$ (red stars). This is not a contradiction of Theorem 2.5 because those components do not lie in the region where $\|T - \hat{T}\| < 10^{-2}$.

CHAPTER 4

TRANSIENT DYNAMICS[1]

4.1 Introduction

Nonlinear differential equations are often used to model economic [BCG09], biological [KS02], chemical [Leh94], and physical [GHA94] systems. Often an equilibrium solution is of interest, and since the equilibrium will not actually be achieved in practice, the behavior of nearby solutions is studied. To make analysis of this problem more tractable, one commonly analyzes stability of the linearized dynamics near the equilibrium. This is done in terms of eigenvalues of some matrix or matrix-valued function. However, linear stability can fail to describe dynamics in practice. If solutions to the linearized system can undergo large transient growth before eventual decay, as can happen for systems $\dot{x} = Mx$ when M is nonnormal [TE05], then the truncated nonlinear terms may become significant and incite even greater growth, rendering the linear stability analysis irrelevant. See [Sin08], [GW06] (the semiconductor laser model which we also study here) and [GG94] for examples where this happens.

Throughout this chapter we will focus on autonomous, homogeneous, constant-coefficient linear systems, which are the type of systems often encountered as linearizations of nonlinear differential equations. Work has already been done on pseudospectral upper and lower bounds on transient dynamics for first-order ODEs of this type (see [TE05]), and those results inspire the bounds derived here. Additionally, upper bounds derived using Lyapunov norms appear in [HP06], and a study of these and more upper bounds, some

[1] This chapter is based on [HB].

elementary and some requiring specific assumptions, is contained in [Pli05]. As for delay differential equations (DDEs), an upper bound has been derived based on Lyapunov-Krasovskii functionals applied to an operator mapping one solution segment to the next [Pli05]; an approximate pseudospectral lower bound is obtained in [GW06] by discretization of the associated infinitesimal generator as in [BM00] to reduce to the ODE case; and in [LBLV92] changes in the time-average of a solution under changes to the model are used to infer effects on transient behavior. As far as the authors are aware, there has been no work extending the pseudospectral bounds in [TE05] to equations beyond first-order ODEs. Our extension consists of replacing the resolvent of a matrix, which plays the key role in the first-order ODE results, with the generalized resolvent of a matrix-valued function which naturally appears in the same way. This idea is straightforward, but a useful implementation depends on the details of the problem at hand. Therefore, rather than state a general result at the cost of introducing an ungainly and narrowly applicable set of assumptions, we apply the main principle to higher order ODEs and somewhat less directly to DDEs with constant delay. We hope to motivate the use of this idea in various other situations, in which the necessary assumptions will be taken into account as needed.

The rest of this chapter is organized as follows. In Section 4.2 we reproduce the transient growth bounds for first-order ODEs and discuss some terminology. In Section 4.3 we make the direct extension to higher-order ODEs and introduce our main example, a model of a semiconductor laser with phase-conjugate feedback. Section 4.4 contains our main theorem, an upper bound for transient growth for DDEs, and its application to a discretized partial DDE and to the laser example. In the penultimate section, we give a practical lower

bound on worst-case transient growth and show its effectiveness on both our examples. Finally, we conclude in Section 4.6.

4.2 Preliminaries

The essential ingredient used in [TE05] to derive bounds on transient growth for ODEs $\dot{x} = Mx$ is the contour integral relationship between the solution propagator e^{tM} and the matrix *resolvent* $(zI - M)^{-1}$. To paraphrase [TE05, Theorem 15.1],

$$(zI - M)^{-1} = \int_0^\infty e^{-zt} e^{tM} \, dt \tag{4.1}$$

for $\mathfrak{Re} z$ sufficiently large, and

$$e^{tM} = \frac{1}{2\pi i} \int_\Gamma e^{zt} (zI - M)^{-1} \, dz \tag{4.2}$$

where Γ is a contour enclosing the eigenvalues of M. The first equation means that $\{\mathcal{L}x\}(z) = (zI - M)^{-1} x_0$ for any solution with initial condition $x(0) = x_0$ and $\mathfrak{Re} z$ sufficiently large, $\mathcal{L}$ being the Laplace transform operator. The second equation is an inverse Laplace transform. By exploiting the Mellin inversion theorem [Wei12], we will be able to use similar integral equations to achieve the desired bounds for more general problems. But first we will collect the necessary terms and state the theorems from [TE05] we wish to extend.

We start by reminding the reader of the definitions of spectrum and pseudospectrum for matrices and matrix-valued functions. The spectrum $\sigma(M)$ of a matrix M is the set of its eigenvalues. The ε-pseudospectrum, denoted by $\sigma_\varepsilon(M)$, is the union of the spectra of all matrices $M + E$, where $\|E\| \leq \varepsilon$. An equivalent definition which gives a different intuition is $\sigma_\varepsilon(M) = \{z \in \mathbb{C} : \|(zI - M)^{-1}\| >$

$\varepsilon^{-1}\}$ [TE05]. Similarly, we defined the spectrum $\Lambda(T)$ and ε-pseudospectrum $\Lambda_\varepsilon(T)$ of a matrix-valued function $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$ in Section 1.2 as the set of eigenvalues of T and the set $\{z \in \Omega : \|T(z)^{-1}\| > \varepsilon^{-1}\}$, respectively. The different symbols σ and Λ will be used to easily distinguish between when we are talking about matrices and when we are talking about matrix-valued functions.

The *spectral abscissa* $\alpha(M)$ of a matrix M is the largest real part among any of its eigenvalues, and the *pseudospectral abscissa* is correspondingly defined as $\alpha_\varepsilon(M) = \max \mathfrak{Re}\sigma_\varepsilon(M)$. The spectral abscissa of M determines asymptotic growth, and the next theorem shows the role of the pseudospectral abscissa in transient growth. Its usefulness is most apparent when $\alpha(M) < 0$.

Theorem 4.1 (based on [TE05], Theorem 15.2)**.** *If M is a matrix and L_ε is the arc length of the boundary of $\sigma_\varepsilon(M)$ (or the convex hull of $\sigma_\varepsilon(M)$) for some $\varepsilon > 0$, then*

$$\|e^{tM}\| \leq \frac{L_\varepsilon e^{t\alpha_\varepsilon(M)}}{2\pi\varepsilon} \qquad \forall t \geq 0. \tag{4.3}$$

Proof. Let Γ be the boundary of $\sigma_\varepsilon(M)$ (or its convex hull). Since $\sigma_\varepsilon(M)$ contains the spectrum of M for every $\varepsilon > 0$, Γ contains the spectrum of M. Therefore we can use the representation (4.2) for e^{tM}. On Γ, $|e^{zt}| \leq e^{t\alpha_\varepsilon(M)}$ and $\|(zI - M)^{-1}\| \leq \varepsilon^{-1}$. Taking norms in (4.2), we then have

$$\|e^{tM}\| \leq \frac{1}{2\pi} e^{t\alpha_\varepsilon(M)} \varepsilon^{-1} \int_\Gamma |dz|,$$

and the theorem follows by observing that $L_\varepsilon = \int_\Gamma |dz|$. □

In addition to upper bounds, pseudospectra also give lower bounds on the maximum achieved by $\|\exp(Mt)\|$, as in this theorem paraphrased from [TE05, Theorem 15.5]:

Theorem 4.2. *Let M be a matrix and let $\omega \in \mathbb{R}$ be fixed. Then $\alpha_\varepsilon(M)$ is finite for each $\varepsilon > 0$ and*

$$\sup_{t\geq 0} \|e^{-\omega t} \exp(tM)\| \geq \frac{\alpha_\varepsilon(M) - \omega}{\varepsilon} \qquad \forall \varepsilon > 0.$$

Proof. Letting $\varepsilon > 0$ be arbitrary, $\alpha_\varepsilon(M)$ is finite because

$$\|(zI - M)^{-1}\| = |z|^{-1} \left\| \sum_{n=0}^{\infty} \left(z^{-1}M\right)^n \right\| \leq \frac{|z|^{-1}}{1 - |z|^{-1}\,\|M\|}$$

is less than ε^{-1} for $|z|$ sufficiently large. Thus, the desired bound is trivially satisfied for $\omega \geq \alpha_\varepsilon(M)$. Therefore we assume that $\omega < \alpha_\varepsilon(M)$.

Now, let $z \in \sigma_\varepsilon(M)$ satisfy $\mathfrak{Re}(z) > \alpha(M)$ so that (4.1) holds, and further suppose that $\mathfrak{Re}(z) > \omega$. Then $\|(zI - M)^{-1}\| \geq \varepsilon^{-1}$ by definition of $\sigma_\varepsilon(M)$, and by (4.1) the hypothesis $\|e^{Mt}\| \leq Ce^{\omega t}$ for all $t \geq 0$ implies that $\|(zI - M)^{-1}\| \leq \frac{C}{\mathfrak{Re}(z)-\omega}$. This in turn implies that $\mathfrak{Re}(z) \leq C\varepsilon + \omega$. Since z may be chosen such that $\mathfrak{Re}(z)$ is arbitrarily close to $\alpha_\varepsilon(M)$, it follows that $\alpha_\varepsilon(M) \leq C\varepsilon + \omega$.

By the contrapositive, if $\alpha_\varepsilon(M) > C\varepsilon + \omega$, then

$$\sup_{t\geq 0} \|e^{-\omega t} e^{Mt}\| \geq \sup \left\{ C : \frac{\alpha_\varepsilon(M) - \omega}{\varepsilon} > C \right\} = \frac{\alpha_\varepsilon(M) - \omega}{\varepsilon}.$$

□

Remark 4.1. If $\omega = 0$ and $\alpha(M) < 0$, then the essence of the theorem is that there is some unit initial condition x_0 such that the solution to $\dot{x} = Mx$, $x(0) = x_0$ satisfies $\|x(t_0)\| \geq \sup_{\varepsilon>0} \frac{\alpha_\varepsilon(M)}{\varepsilon}$ at some finite time t_0 before eventually decaying to zero. One can think of such a solution as a long-lived "pseudo-mode" associated with pseudo-eigenvalues in the right half-plane. If the ε-pseudospectrum extends far into the right half-plane for some small ε, then there must be some solution that exhibits large transient growth.

To bound transient growth for the higher-order ODE and DDE cases, a *generalized resolvent* $T(z)^{-1}$, with $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{n\times n}$ a matrix-valued function, will play the role that the resolvent $(zI - M)^{-1}$ did above. We let $\alpha(T)$ represent the largest real part of any eigenvalue of T and call it the *spectral abscissa of* T. Similarly, we define the *pseudospectral abscissa of* T as

$$\alpha_\varepsilon(T) = \sup_{z\in\Lambda_\varepsilon(T)} \Re\mathrm{e} z.$$

The following proposition is immediate.

Proposition 4.1. *If* $\|T(z)^{-1}\| \to 0$ *uniformly as* $\Re\mathrm{e} z \to \infty$, *then* $\alpha_\varepsilon(T) < \infty$ *for all* $\varepsilon > 0$.

4.3 Upper bounds for higher-order ODEs

In this section we treat equations of the form

$$y^{(n)} = \sum_{j=0}^{n-1} A_j y^{(j)} \tag{4.4}$$

with initial conditions $y^{(j)}(0) = y_0^{(j)}$, $j = 0, ..., n-1$, and where each $A_j \in \mathbb{C}^{k\times k}$.

We can solve (4.4) by writing it in first-order form, e.g., $\dot{x} = Mx$, $x = [y, \dot{y}, ..., y^{(n-1)}]^T$, where

$$M = \begin{bmatrix} 0 & I & 0 & \dots & 0 \\ 0 & 0 & I & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & I \\ A_0 & A_1 & A_2 & \dots & A_{n-1} \end{bmatrix}. \tag{4.5}$$

Then Theorem 4.1 can be applied to the solutions $x(t) = e^{Mt}x(0)$ in order to bound $y(t)$, since $\|y(t)\| \le \|x(t)\|$. But since the maximum reached by $\|x(t)\|$ could be much

larger than the maximum of $\|y(t)\|$, one can hardly expect to obtain a tight bound for $\|y(t)\|$ with this process. With the next theorem, we can bound $y(t)$ directly.

Theorem 4.3. *Let the equations in y and x be as above, with M partitioned as*

$$M = \begin{bmatrix} 0 & B \\ C & D \end{bmatrix}, \qquad D \text{ square.}$$

Assume $y_0^{(j)} = 0$ *for* $j = 1, ..., n-1$. *Then* $y(t) = \Psi(t)y(0)$ *with*

$$\|\Psi(t)\| \leq \frac{L_\varepsilon e^{\alpha_\varepsilon(T)t}}{2\pi\varepsilon} \qquad \forall \varepsilon > 0,$$

where $T(z) = zI - B(zI - D)^{-1}C$ *and* L_ε *is the arc length of the* ε*-pseudospectrum of* T. *The bound is finite for every* $\varepsilon > 0$.

Proof. Let E_1 represent the first k columns of the $nk \times nk$ identity. Then the initial condition in x is $x(0) = E_1 y(0)$, so that $x(t) = e^{Mt}E_1 y(0)$ and hence $y(t) = E_1^T e^{Mt} E_1 y(0)$. Therefore we define $\Psi(t) := E_1^T e^{Mt} E_1$. From the integral representation (4.2) for e^{Mt}, we have

$$\Psi(t) = \frac{1}{2\pi i} \int_\Gamma \underbrace{E_1^T (zI - M)^{-1} E_1}_{T(z)^{-1}} e^{zt}\, dz.$$

As in the proof of Theorem 4.2, for $|z|$ large enough $\|T(z)^{-1}\| \leq \varepsilon^{-1}$. Therefore $\Lambda_\varepsilon(T)$ is bounded for every $\varepsilon > 0$. Therefore L_ε and $\alpha_\varepsilon(T)$ are both finite and the result follows as in Theorem 4.1. □

Remark 4.2. Bounds for similar objects can be found in [Pli05] in the section "Kreiss Matrix and Hille-Yosida Generation Theorems," where structured (M, β)-stability is considered.

The assumption $y^{(j)}(0) = 0$ for $j = 1, ..., n-1$ was not essential, as the following corollary shows.

Corollary 4.1. *If y satisfies $y^{(n)} = \sum_{j=0}^{n-1} A_j y^{(j)}$, with initial condition $y^{(j)}(0) = y_0^{(j)}$ for $j = 0, 1, ..., n-1$, then*

$$\|y(t)\| \leq \sum_{j=0}^{n-1} \frac{L_\varepsilon^{(j)} e^{\alpha_\varepsilon(T_j)t}}{2\pi\varepsilon} \|y_0^{(j)}\|, \qquad \forall \varepsilon > 0,$$

where $L_\varepsilon^{(j)}$ is the arclength of the boundary of $\Lambda_\varepsilon(T_j)$, $T_j(z)^{-1} = E_1^T(zI - M)^{-1}E_{j+1}$, and E_{j+1} is the $j+1$-th block column of the $nk \times nk$ identity partitioned into n block columns.

Proof. From $x(0) = \sum_{j=0}^{n-1} E_{j+1} y_0^{(j)}$, we can use (4.2) to write

$$y(t) = \sum_{j=0}^{n-1} \frac{1}{2\pi i} \int_\Gamma T_j(z)^{-1} e^{zt}\, dz \cdot y^{(j)}(0), \qquad T_j(z)^{-1} = E_1^T(zI - M)^{-1}E_{j+1}$$

and apply the theorem to each summand. □

Remark 4.3. We arrive at expressions for each $T_j(z)$ by taking the Laplace transform of the original equation (4.4) and expressing $y(t)$ in terms of the inverse Laplace transform. First, using standard facts about the Laplace transform,

$$s^n Y(s) - \sum_{j=0}^{n-1} s^{n-1-j} y_0^{(j)} = \sum_{k=0}^{n-1} A_k \left(s^k Y(s) - \sum_{j=0}^{k-1} s^{k-1-j} y_0^{(j)} \right), \qquad Y = \mathcal{L}y.$$

Rearranging,

$$\underbrace{\left(s^n I - \sum_{k=0}^{n-1} s^k A_k \right)}_{P(s)} Y(s) = \sum_{j=0}^{n-1} \underbrace{\left(s^{n-1-j} I - \sum_{k=j+1}^{n-1} A_k s^{k-1-j} \right)}_{X_j(s)} y_0^{(j)}.$$

Then we recover

$$y(t) = \sum_{j=0}^{n-1} \frac{1}{2\pi i} \int_{\Gamma_j} P(z)^{-1} X_j(z)\, e^{zt}\, dz \cdot y_0^{(j)}$$

from which we see that $T_j(z)^{-1} = P(z)^{-1} X_j(z)$. Therefore $T_j(z) = X_j(z)^{-1} P(z)$.

The last result of this section is the higher-order difference equation version of the last corollary, and is a direct extension of [TE05, Theorem 16.2].

Corollary 4.2. *Suppose* (y_n) *satisfies the difference equation* $y_{n+1} = \sum_{j=0}^{N} A_j y_{n-j}$ *with initial conditions* $y_0, y_{-1}, ..., y_{-N}$ *given. Then*

$$\|y_n\| \leq \sum_{j=0}^{N} \frac{L_\varepsilon^{(j)} \rho_\varepsilon(T_j)^n}{2\pi\varepsilon} \|y_{0-j}\| < \infty \qquad \forall \varepsilon > 0$$

where $T_j(z)^{-1} = E_1^T (zI - M)^{-1} E_{j+1}$, $L_\varepsilon^{(j)}$ *is the arclength of the boundary of* $\Lambda_\varepsilon(T_j)$, *and* $\rho_\varepsilon(T_j) = \max\{|z| : z \in \Lambda_\varepsilon(T_j)\}$ *is the pseudospectral radius.*

Proof. Putting $x_n = [y_{n-0}, y_{n-1}, ..., y_{n-N}]^T$, we have $x_n = M^n x_0$. Applying the inverse Z-transform to $z(zI - M)^{-1}$ we obtain $M^n = \frac{1}{2\pi i} \int_\Gamma z^n (zI - M)^{-1}\, dz$ for Γ a contour enclosing the spectrum of M. The quantity of interest may then be expressed as $y_n = E_1^T x_n = \sum_{j=0}^{N} E_1^T M^n E_{j+1} y_{0-j}$. Since $E_1^T M^n E_{j+1} = \frac{1}{2\pi i} \int_{\Gamma_j} z^n T_j(z)^{-1}\, dz$, the bound for this term follows by taking Γ_j equal to the ε-pseudospectrum of T_j. These bounds are finite for any ε since $\|(zI - M)^{-1}\| \geq \|T_j(z)^{-1}\|$ for all z implies that $\Lambda_\varepsilon(T) \subset \Lambda_\varepsilon(M)$, and we know the latter to be finite. □

Our first example demonstrates the improvement in bounding the solution to (4.4) directly versus bounding the solution to the first-order form $\dot{x} = Mx$ while simultaneously motivating the need for the bound in the next section.

Example 4.1. The model for a semiconductor laser with phase-conjugate feedback studied in [GW06] has an equilibrium at

$$(E_x, E_y, N) = (+1.8458171368652383, -0.2415616277234652, +7.6430064479131916)$$

after scaling, and linearizing about this equilibrium yields the DDE $\dot{y}(t) = Ay(t) +$

$By(t-1)$ where[2]

$$A = \begin{bmatrix} -8.4983 \times 10^{-1} & 1.4786 \times 10^{-1} & 4.4381 \times 10^{1} \\ 3.7540 \times 10^{-3} & -2.8049 \times 10^{-1} & -2.2922 \times 10^{2} \\ -1.7537 \times 10^{-1} & 2.2951 \times 10^{-2} & -3.6079 \times 10^{-1} \end{bmatrix},$$

$$B = \begin{bmatrix} 2.8000 \times 10^{-1} & 0 & 0 \\ 0 & -2.8000 \times 10^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Discretizing with $N+1$ points on each unit segment, we can use the forward Euler approximation to obtain the higher-order difference equation approximation $y_{j+1} = (I+hA)y_j + hBy_{j-N}$, $h = 1/N$. Initial conditions $y_0, y_{0-1}, ..., y_{0-N}$ come from sampling the initial condition for the original equation on $[-1, 0]$. We then apply Corollary 4.2 with $A_N = hB$, $A_0 = I + hA$, and $A_j = 0$ otherwise. A companion linearization gives

$$\underbrace{\begin{bmatrix} y_{j+1} \\ \vdots \\ y_{j-N+1} \end{bmatrix}}_{x_{n+1}} = \underbrace{\begin{bmatrix} I + hA & 0 & \dots & hB \\ I & & & 0 \\ & \ddots & & \vdots \\ & & I & 0 \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} y_j \\ \vdots \\ y_{j-N} \end{bmatrix}}_{x_n}.$$

Here we choose the initial condition $y(t) = 0.0015 \times (Ex, Ey, N)^T$ on $[-1, 0]$ since a similar initial condition in [GW06] corresponds to a decaying solution with

[2] The equilibrium and linearized system computed here differ slightly from those in [GW06]. It appears there were two typographical errors and a lack of precision in one or more of the given parameters, which led to the parameters, linearization, and equilibrium stated in [GW06] being mutually inconsistent. The authors have been contacted, and we have attempted to reproduce their linearization and equilibrium closely by making the following adjustments. We have put $I = 0.0651A$ to match the value in [GHA94], [GK02], and [KGL98] cited in [GW06] as the sources for the parameters. We have set $\kappa = 4.2 \times 10^8 s^{-1}$ so that the coefficient matrix B for the delay term is the same as in [GW06]. Lastly, we have put $N_{sol} = N_0 + 1/(G_N \tau_p)$ as prescribed in [GK02] (which is approximately the value of N_{sol} stated in [GW06]). The equilibrium we have listed above was computed using Newton's method after the parameter corrections were implemented, with the equilibrium stated in [GW06] as an initial guess.

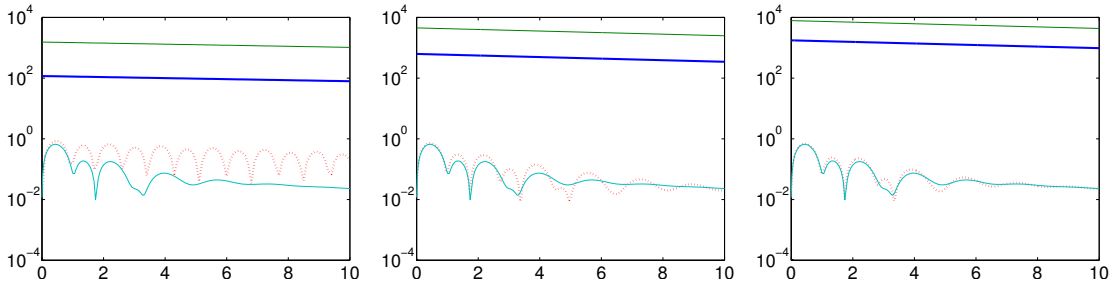

Figure 4.1: Upper bound from Corollary 4.2 (thick, solid), upper bound from Theorem 4.1, 2-norm of solution $y(t)$ to continuous DDE computed with the MATLAB `dde23` routine, and 2-norm of solution y_n to discretized equation (thick, dotted) for $N = 10$ (left), $N = 25$ (center), $N = 50$ (right).

nontrivial transient growth.

As in the continuous case, with a little manipulation we find that $T_j(z)^{-1} = P(z)^{-1}X_j(z)$, where $P(z) = z^{N+1}I - (I + hA)z^N - hB$, $X_0(z) = z^N I$, and $X_j(z) = hBz^{j-1}$ for $j \geq 1$. (In general, $P(z) = z^{N+1}I - \sum_{j=0}^{N} A_j z^{N-j}$ and $X_j(z) = \sum_{k=j}^{N} A_k z^{N+j-1-k}$ for $j \geq 1$.) Notice that B is singular and therefore the inverse of $T_j(z)^{-1}$ does not exist for $j \geq 1$. However, we can still do $\|T_j(z)^{-1}\| \leq \|P(z)^{-1}\|\|h\|\|B\|\,|z^{j-1}|$ and obtain bounds by taking $\Gamma_j = \partial\{z : \|P(z)^{-1}\|\|h\|\|B\|\,|z|^{j-1} = \varepsilon^{-1}\}$ for $j \geq 1$. In Figure 4.1, we show an upper bound on $\|y_n\|_2$ from using Corollary 4.2, an upper bound on $\|x_n\|_2$ using Theorem 4.1, the 2-norm of the solution y_n and the 2-norm of the solution $y(t)$ to the continuous DDE. Notice that as the mesh becomes finer, y_n becomes a better approximation to $y(t)$ but the upper bound on $\|y_n\|_2$ becomes much more generous. This is because the spectral radius of M increases with mesh size. This in turn suggests that a bound which comes directly from the continuous DDE itself may be more straightforward and effective.

4.4 Upper bounds for delay differential equations

Now we turn to transient bounds for DDEs

$$\dot{u}(t) = Au(t) + Bu(t-\tau) \qquad A, B \in \mathbb{C}^{n\times n} \tag{4.6}$$

with a single delay $\tau > 0$ and with $\alpha(A)$ and $\alpha(T)$ both negative, where $T(z) = zI - A - Be^{-\tau z}$ is the characteristic equation [MN07b]. Although we treat only a single delay here, a direct extension to multiple constant delays is straightforward.

The characteristic equation $T(z) = zI - A - Be^{-\tau z}$ generally has infinitely many eigenvalues, as is often the case with nonlinear matrix-valued functions. Therefore, unlike in the previous section, bounds on transient behavior will depend on integrals whose integration path is unbounded. So, if we expect a bound for a DDE to be useful in practice, then we expect it to require more preparation (such as locating eigenvalues so as to find an admissible integration path) and look more complicated (since the integrand's behavior at infinity will need to be analyzed) than a bound for an ODE.

Let Ψ be the fundamental solution for the DDE, that is, the solution whose initial conditions are zero on $[-\tau, 0)$ and the matrix identity I at $t = 0$. Following the treatment in Chapter 1 of [HL93], we first bound Ψ by invoking the Mellin inversion theorem [Wei12] and then splitting the characteristic equation into its linear and nonlinear parts. We show that the integration can be taken over a curve more convenient than the usual vertical one, and finally compute upper bounds using elementary means.

Lemma 4.1. *If $X, Y \in \mathbb{C}^{n\times n}$ are two matrices, and if $\|X^{-1}\|\,\|Y\| < 1$, then*

$$\|(X-Y)^{-1}\| \le \frac{1}{\|X^{-1}\|^{-1} - \|Y\|}.$$

Proof. We write $(X - Y)^{-1} = (I - X^{-1}Y)^{-1}X^{-1}$. By hypothesis, the Neumann series $\sum_{j=0}^{\infty}(X^{-1}Y)^j$ for $(I - X^{-1}Y)^{-1}$ converges and is bounded by $(1 - \|X^{-1}\|\,\|Y\|)^{-1}$. Therefore $\|(X - Y)^{-1}\| \le \|X^{-1}\|\,(1 - \|X^{-1}\|\,\|Y\|)^{-1}$ and the desired result follows. □

Lemma 4.1 can be used to bound $\|T(z)^{-1}\| = \|(zI - A - Be^{-\tau z})^{-1}\|$ in various ways, depending on the choice of norm and the properties of the matrices A and B. For simplicity, in the following lemma we use $\|\cdot\| = \|\cdot\|_2$ and assume A is Hermitian. Generalizations are straightforward. For instance, the laser example analyzed in this section does not have A Hermitian and we show how to apply our results to that case.

Lemma 4.2. *Let $T(z) = zI - A - Be^{-\tau z}$ with $A = VDV^*$ Hermitian. If y_0 is a given positive number, and η is chosen so that*

$$1 < \eta y_0 < \min\left\{\frac{y_0}{\|B\|_2}, e^{-\alpha(T)\tau}, e^{-\alpha(A)\tau}\right\},$$

then

$$\Gamma = \underbrace{\{x(y) + iy : |y| > y_0\}}_{\Gamma_\infty} \cup \underbrace{\{x_0 + iy : |y| \le y_0, x_0 = x(y_0)\}}_{\Gamma_0}, \qquad x(y) = -\frac{1}{\tau}\log(|y|\eta)$$

is to the right of both $\Lambda(T)$ and $\sigma(A)$ but lies entirely in the left half-plane.

Proof. Γ is certainly to the right of $\sigma(A)$, since all eigenvalues of A are real and the condition $\eta y_0 < e^{-\alpha(A)\tau}$ guarantees $x(y_0) > \alpha(A)$. The eigenvalues of T are also to the left of Γ_0 by the condition $\eta y_0 < e^{-\alpha(T)\tau}$. As for Γ_∞, $\|(zI - A)^{-1}\|_2^{-1} = \sigma_{\min}(zI - D) \ge |y|$ because the eigenvalues of A are real. Hence, if $z \in \Gamma_\infty$, then $\|(zI - A)^{-1}\|_2^{-1} \ge |y| > \|B\|_2\eta|y|$ by the hypothesis $\eta < 1/\|B\|_2$. Therefore, Lemma 4.1 applies with $X = zI - A$ and $Y = Be^{-\tau z}$, so $\|T(z)^{-1}\| \le (|y| - \|B\|_2\eta|y|)^{-1} < \infty$ on Γ_∞. Since decreasing η to zero moves Γ_∞ infinitely to the right, and decreasing η does not violate the assumption guaranteeing nonsingularity of T on Γ_∞, it follows

that T is nonsingular on and at all points to the right of Γ_∞. Therefore Γ is to the right of $\Lambda(T)$. The condition $1 < \eta y_0$ assures that Γ_0 is in the left half-plane, and therefore so is Γ. □

By our assumption that all eigenvalues of T are in the left half-plane, all solutions of (4.6) are exponentially stable [MN07b, Proposition 1.6] and hence of exponential order. Therefore we can take the Laplace transform of (4.6) to obtain

$$(zI - A - Be^{-\tau z})U(z) = u(0) + Be^{-\tau z}\int_{-\tau}^{0} e^{-zt}u(t)\,dt, \quad U = \mathcal{L}u.$$

Since the fundamental solution satisfies $\Psi(t) = 0$ on $[-\tau, 0)$ and $\Psi(0) = I$, it follows that $(\mathcal{L}\Psi)(z) = T(z)^{-1}$. Then we can use the Mellin inversion theorem [Wei12] to write

$$\Psi(t) = \frac{1}{2\pi i}\int_{\gamma+i\mathbb{R}} T(z)^{-1}e^{zt}\,dz \tag{4.7}$$

for any $\gamma > \alpha(T)$. The next lemma shows that we can integrate over the contour Γ from Lemma 4.2 rather than $\gamma + i\mathbb{R}$ in (4.7).

Lemma 4.3. *For Γ as in Lemma 4.2 and γ such that (4.7) holds, we have*

$$\int_{\Gamma} T(z)^{-1}e^{zt}\,dz = \int_{\gamma+i\mathbb{R}} T(z)^{-1}e^{zt}\,dz.$$

Proof. Since T has no eigenvalues in the region bounded by $\gamma + i\mathbb{R}$ and Γ, we only need to show that the integrals

$$\int_{x(y)}^{\gamma} T(w + iy)^{-1}\, e^{(w+iy)t}\,dw$$

go to zero as $y \to \pm\infty$. But from Lemma 4.1 we know $\|T(w + iy)^{-1}\| \sim \frac{1}{|y|}$ on $x(y) \le w \le \gamma$ as $|y|$ becomes large, and $|e^{(w+iy)t}| \le e^{\gamma t}$ on the integration path which itself has arc length $\sim \log(|y|)$. Therefore

$$\left\| \int_{x(y)}^{\gamma} T(w + iy)^{-1}\, e^{(w+iy)t}\,dw \right\| \lesssim \frac{\log|y|}{|y|} \to 0$$

as $|y|$ becomes large, and the lemma is proved. □

We now come to the main result of this section, in which we bound transient growth of the fundamental solution. Note that a bound on $\Psi(t)$ for $t \geq \tau$ is all that is required, since $\Psi(t) = e^{At}$ for $0 \leq t < \tau$. Again, we use the 2-norm, but only for simplicity.

Theorem 4.4. *With the hypotheses of the previous lemmas, the fundamental solution of $\dot{u}(t) = Au(t) + Bu(t-\tau)$ satisfies the bound*

$$\|\Psi(t)\|_2 \leq \|\exp(At)\|_2 + e^{x_0 t} I_0 + e^{x_0 t} \frac{C}{t/\tau}$$

on $t \geq \tau$, where

$$I_0 = \frac{1}{2\pi} \int_{-y_0}^{y_0} \|T(x_0 + iy)^{-1} - R(x_0 + iy)\|_2 \, dy, \quad R(z) = (zI - A)^{-1}$$

and

$$C = \frac{\|B\|_2 \eta \sqrt{(\tau y_0)^{-2} + 1}}{\pi (1 - \|B\|_2 \eta)}.$$

Proof. Since Γ was chosen to the right of all eigenvalues of A, the splitting

$$\frac{1}{2\pi i} \int_\Gamma T(z)^{-1} e^{zt} \, dz = \frac{1}{2\pi i} \int_\Gamma R(z) e^{zt} \, dz + \frac{1}{2\pi i} \int_\Gamma \left[T(z)^{-1} - R(z) \right] e^{zt} \, dz$$

and subsequent evaluation of the first summand as e^{At} is justified, as the Mellin inversion theorem applies to $R(z)$ for the same reason it applies to $T(z)^{-1}$. With I_0 as defined in the theorem statement, it only remains to give a bound on the second integral in the sum.

From the hypothesis that A is Hermitian we have that $\|R(z)\|_2 \leq |y|^{-1}$, and hence $\|R(z)Be^{-\tau z}\|_2 \leq \|B\|_2 \eta$. Therefore the assumption $\eta y_0 < y_0 / \|B\|_2$ implies $\|R(z)Be^{-\tau z}\|_2 < 1$ on Γ_∞, so that $T(z)^{-1} - R(z)$ is subject to the Neumann bound

$$\|T(z)^{-1} - R(z)\|_2 \leq |y|^{-1} \|B\|_2 \eta \, (1 - \|B\|_2 \eta)^{-1} .$$

In addition, if $z \in \Gamma_\infty$ then $|z'(y)| \leq \sqrt{(\tau y_0)^{-2} + 1}$. It then follows that

$$\left\| \frac{1}{2\pi i} \int_{\Gamma_\infty} \left[T(z)^{-1} - R(z) \right] e^{zt} \, dz \right\|_2 \leq e^{x_0 t} \frac{C}{t/\tau}.$$

□

Remark 4.4. In general, if A is not Hermitian we can still bound $\|R(z)\|$ simply by splitting A into its Hermitian and skew-Hermitian parts as $A = H + S$, from which $\|R(z)\|_2 \leq (|y| - \|S\|_2)^{-1}$ if we use the 2-norm. However $\|R(z)\|$ is bounded must be taken into account when choosing Γ_∞.

Also note that we could have integrated over a vertical contour at $\Re\mathrm{e} z = x_0$, because $\|T(x_0 + iy)^{-1} - R(x_0 + iy)\| \lesssim |y|^{-2}$ as $|y| \to \infty$. But then we obtain

$$\|\Psi(t)\|_2 \leq \|\exp(At)\|_2 + e^{x_0 t} I_0 + e^{x_0 t} C, \qquad C = \frac{\|B\|_2 \eta}{\pi(1 - \|B\|_2 \eta)} \tag{4.8}$$

and we have lost the $1/t$ dependence in the third term.

Since $\|T(z)^{-1}\|$ is integrable on Γ_∞, we also could have obtained an upper bound in terms of the integral of $\|T(z)^{-1}\|$ over Γ_0 and another term with $1/t$ dependence, specifically

$$\|\Psi(t)\|_2 \leq e^{x_0 t} \tilde{I}_0 + e^{x_0 t} \frac{C}{t/\tau}, \qquad \tilde{I}_0 = \frac{1}{2\pi} \int_{\Gamma_0} \|T(z)^{-1}\|_2 \, |dz|, \;\; C = \frac{\sqrt{(\tau y_0)^{-2} + 1}}{\pi(1 - \|B\|_2 \eta)}. \tag{4.9}$$

In this version we fail to take advantage of the closed form of $\int R(z) e^{zt} \, dz$. However, we may be able to shift x_0 further to the left since we no longer need to have Γ to the right of the spectrum of A, and this will result in faster decay.

One consequence of a bound on the fundamental solution is a bound on worst-case transient behavior of a certain class of solutions, namely the ones equal to zero on $[-\tau, 0)$ and with an initial "shock" condition specified at $t = 0$.

Example 4.2. Consider the DDE

$$\dot{v}(t) = Av(t) + Bv(t-\tau), \quad \tau = 0.2 \tag{4.10}$$

coming from the discretization of a parabolic partial delay differential equation[3] where $n = 10^2$ and $h = \pi/(n+1)$, and $A, B \in \mathbb{C}^{n,n}$ are defined by

$$A = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & 1 & \\ & 1 & -2 \end{pmatrix} + 2I, \qquad B(j,j) = a_1(jh), \quad a_1(x) = -4.1 + x(1 - e^{x-\pi}).$$

Rather than compute eigenvalues of T to obtain $\alpha(T)$, we can compute inclusion regions by applying Theorem 2.1.

For $AV = \Lambda V$ an eigendecomposition of A, the matrix-valued function $\tilde{T}(z) = zI - \Lambda - V^{-1}BVe^{-\tau z}$ has the same spectrum as T. Applying Theorem 2.1 to $\tilde{T}$ with a diagonal/off-diagonal splitting yields the inclusion regions plotted in Figure 4.2 (left). The rightmost point of the inclusion regions is then a bound for $\alpha(T)$. The largest eigenvalues of A are also plotted, and the solid contour is chosen as in the theorem, with $y_0 = 21.4214$ and $\eta = 0.0491366$, so that it is to the right of both the eigenvalues of T and the eigenvalues of A. The dashed contour is the vertical alternate integration path, as referred to in Remark 4.4. Lastly, if we use (4.9), we can integrate over a contour whose vertical section is shifted to the left, depicted as the dotted line in Figure 4.2 (left). In Figure 4.2 (right), we have plotted the bound derived using the theorem (solid), as well alternate bounds (4.8) (dashed), (4.9) (dot-dashed), and (4.9) with the contour whose vertical part is shifted leftward (dotted). The last bound gives better results for larger times, as expected, but the bound given in Theorem 4.4 outperforms it for smaller times and outperforms the other bounds for all times $t > \tau$.

[3] This example is adapted from the example in [Eff13, §5.1]. We treated the latter in Chapter 3 where it is named the single delay PDDE and we use the same localization techniques here.

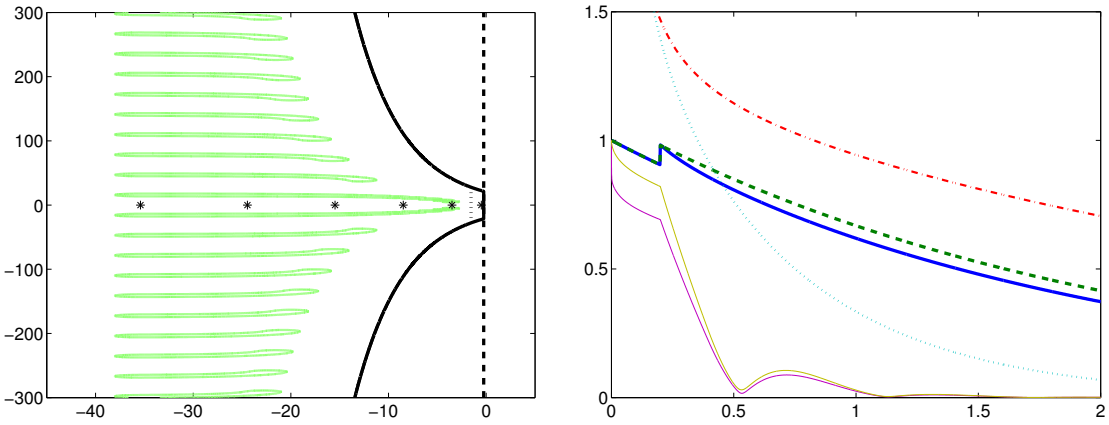

Figure 4.2: Left: Inclusion regions for the eigenvalues of $T(z)$ (green line), the six largest eigenvalues of A ($*$), the contour Γ used in Theorem 4.4 (solid), and the vertical contour used for alternate bound (4.8) (dashed). Contour parameters y_0 and η are set to 21.4214 and 0.0491366, respectively. For the contour with left-shifted vertical part (dotted), $y_0 = 27.7628$ and $\eta = 0.0491366$. Right: Upper bounds on the fundamental solution of (4.10): the upper bound described in the theorem (solid), and the alternates (4.8) (dashed), (4.9) (dot-dashed), and (4.9) with the contour with dotted vertical part (dotted). Lower solid curves are solutions to (4.10) with zero initial condition on $[-\tau, 0)$ and various unit-norm initial conditions specified at $t = 0$.

In case $u([-\tau, 0)) \not\equiv 0$, we can still obtain bounds on $u(t)$ in terms of the fundamental solution $\Psi(t)$.

Corollary 4.3. *Suppose u satisfies the DDE of the theorem subject to initial conditions $u(0+) = u_0$ and $u(t) = \varphi(t)$ on $[-\tau, 0)$, with $B\varphi$ integrable. Then*

$$\|u(t)\|_2 \le \|\Psi(t)\|_2 \cdot \|u_0\|_2 + \sup_{0\le\nu<\tau} \|\Psi(t-\nu)\|_2 \int_0^\tau \|B\varphi(\nu-\tau)\|_2 \, d\nu.$$

Furthermore, this bound is dominated by the more generous but more explicit piecewise bound $k_1(t)\|u_0\|_2 + k_2(t)\int_0^\tau \|B\varphi(\nu-\tau)\|_2 \, d\nu$, where $k_1(t)$ is an upper bound on $\|\Psi(t)\|_2$ and

$$k_2(t) = \begin{cases} \sup\limits_{0<s<t} \|\exp(As)\|_2 & (0 \le t < \tau), \\ \sup\limits_{t-\tau<s<\tau} \|\exp(As)\|_2 + \sup\limits_{\tau<s<t} \|\exp(As)\|_2 + e^{x_0\tau}(I_0 + C) & (\tau \le t < 2\tau), \\ \sup\limits_{t-\tau<s<t} \|\exp(As)\|_2 + e^{x_0(t-\tau)}\left(I_0 + \dfrac{C}{(t-\tau)/\tau}\right) & (t \ge 2\tau). \end{cases}$$

Proof. From [HL93, Ch. 1, Thm. 6.1],

$$u(t) = \Psi(t)u_0 + \int_0^\tau \Psi(t-\nu)B\varphi(\nu-\tau)\,d\nu$$

and the first assertion is obvious. Using the fact that $\Psi(t) = e^{At}$ on $0 \le t < \tau$ and $\Psi(t-\nu) = 0$ for $t - \nu < 0$, the integration path can be truncated to $\int_0^t$. Similarly, if $\tau \le t < 2\tau$, then $\Psi(t-\nu) = e^{A(t-\nu)}$ for $t - \tau < \nu < \tau$, and for this range of ν we have $t - \nu \ge \tau$. Finally, if $t \ge n\tau$ then $t - \nu \ge (n-1)\tau$ as $\nu \le \tau$. Taking norms and applying the theorem gives the piecewise bounds. □

Example 4.3. We return to the model of the semiconductor laser with phase-conjugate feedback. Since A is not Hermitian, Theorem 4.4 does not apply directly. However, by letting β equal the largest imaginary part of any eigenvalue of A, changing the definition of Γ_∞ so that $x(y) = -\log(\eta(|y| - \beta))$ instead, and using the fact that $A = VDV^{-1}$ is diagonalizable, it is straightforward to derive the same bound as in Theorem 4.4 with the alteration

$$C = \frac{\kappa_2(V)\eta\|E\|_2\sqrt{(\tau(y_0-\beta))^{-1}+1}}{\pi(1-\eta\|E\|_2)},$$

where $E = V^{-1}BV$, $\kappa_2(V)$ is the 2-norm condition number of V, and y_0 and η were chosen to satisfy

$$1 < \eta(y_0 - \beta) < \min\left\{\frac{y_0-\beta}{\|E\|_2}, e^{-\alpha(T)}, e^{-\alpha(A)}\right\}.$$

Inclusion regions were obtained for $\tilde{T}(z) = zI - \Lambda - Ee^{-z}$ with the splitting $D(z) = zI - \Lambda - E_0e^{-z}$ and $E(z) = Fe^{-z}$ ($E_0 = \operatorname{diag} E$, $F = E - E_0$) and an application of Theorem 2.1. The one component of the inclusion region intersecting the right half-plane contains exactly one eigenvalue of $D(z)$, and therefore exactly one eigenvalue of $T(z)$, which we have computed using Newton's method on a bordered system [Gov00, Chapter 3].

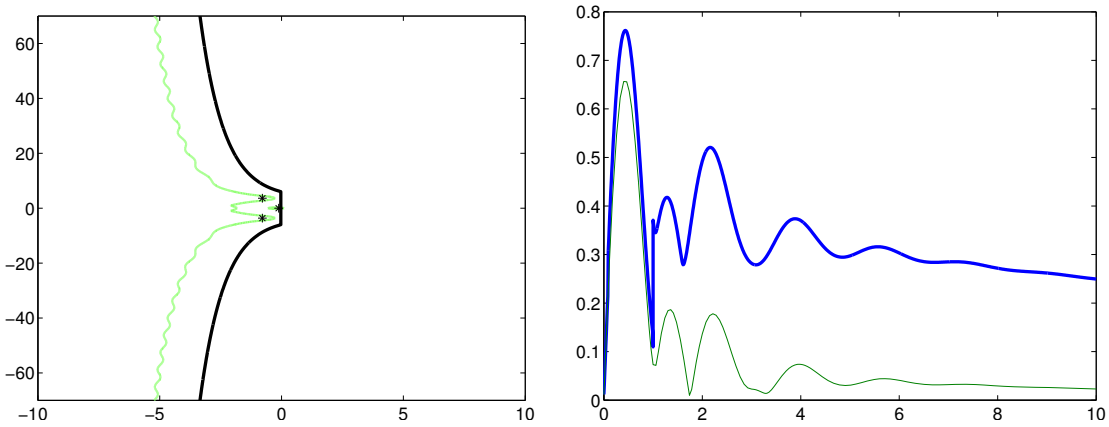

Figure 4.3: Left: An inclusion region for the eigenvalues of $T(z)$, the three rightmost eigenvalues of $T(z)$, and the contour Γ (thick line). Contour parameters y_0 and η are set to 21.4214 and 0.0491366, respectively. Right: Upper bound on solution with initial conditions $\phi(t) \equiv u(0)$ (thick line) and the solution computed with MATLAB's `dde23`.

4.5 Lower bounds

The following is an extension of Theorem 4.2.

Theorem 4.5. *Let T be a matrix-valued function and suppose we have the representations*

$$\Psi(t) = \frac{1}{2\pi i}\int_\Gamma T(z)^{-1}e^{zt}\,dz \qquad \text{and} \qquad T(z)^{-1} = \int_0^\infty \Psi(t)e^{-zt}\,dt$$

where the latter holds for $\Re\mathrm{e}(z) > \alpha(T)$ and Γ is a (possibly unbounded) curve in the complex plane. Then for arbitrary $\omega \in \mathbb{R}$,

$$\sup_{t\geq 0}\|\Psi(t)\|e^{-\omega t} \geq \frac{\alpha_\varepsilon(T)-\omega}{\varepsilon}$$

for any $\varepsilon > 0$ for which $\alpha_\varepsilon(T)$ is finite.

Proof. Suppose $\|\Psi(t)\| \leq Ce^{\omega t}$ for all $t \geq 0$ and fix $\varepsilon > 0$. Without loss of generality, suppose that $\omega < \alpha_\varepsilon(T)$, and take $z \in \Lambda_\varepsilon(T)$ such that $\Re z > \alpha(T)$ and $\Re z > \omega$. Then by the representation for $T(z)^{-1}$, the bound on $\Psi(t)$ implies $\|T(z)^{-1}\| \leq$

$\frac{C}{\Re\mathfrak{e}(z)-\omega}$. Then $\|T(z)^{-1}\| > \varepsilon^{-1}$ by definition of $\Lambda_\varepsilon(T)$, which implies $\Re\mathfrak{e}(z) - \omega < C\varepsilon$. It follows that $\alpha_\varepsilon(T) \leq C\varepsilon + \omega$. By the contrapositive, $\alpha_\varepsilon(T) > C\varepsilon + \omega$ implies the desired result. □

The following proposition is easier to use in practice.

Proposition 4.2. *For each $x > 0$, $\sup_{t\geq 0} \|\Psi(t)\| \geq x \sup_{y\in\mathbb{R}} \|T(x + iy)^{-1}\|$.*

Proof. Fix x and let $y \in \mathbb{R}$ be arbitrary. Set $\|T(x+iy)^{-1}\| = \varepsilon^{-1}$. Then $\alpha_\varepsilon(T) \geq x$. From Theorem 4.5, $\sup_{t\geq 0} \|\Psi(t)\| \geq x\varepsilon^{-1}$. For fixed x, the right-hand side is maximized by finding ε as small as possible, i.e., by finding y such that $\|T(x + iy)^{-1}\|$ is as large as possible. □

Remark 4.5. Note that for a given x we may only need to check a finite range of y values. This can be shown by proving that for $|y|$ sufficiently large $\|T(x+iy)^{-1}\| \leq \|T(x)^{-1}\|$, for example.

Example 4.4. We now give lower bounds on worst-case growth for our two examples. In the case of the discretized PDDE, we use the fact that A is Hermitian to derive $\|T(x + iy)^{-1}\|_2 \leq \|T(x)^{-1}\|_2$ for $|y| > |x| + \|B\|_2 e^{-\tau x} + \sigma_{\min}(T(x))$ as per the previous remark, and check 100 equally spaced x values in $[5, 10]$ for the largest lower bound given by Proposition 4.2. For the linearization of the laser example, where A is not Hermitian, we use instead that $\|T(x + iy)^{-1}\|_2 \leq \|T(x)^{-1}\|_2$ for $|y| > \|xI - A\|_2 + \|B\|_2 e^{-x} + \sigma_{\min}(T(x))$ and check an equally spaced 100 point mesh of $[1, 5]$.

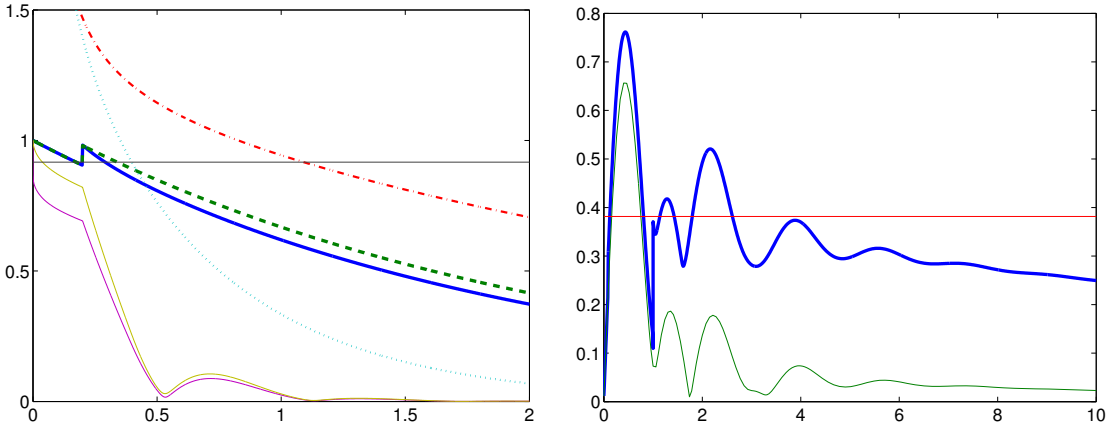

Figure 4.4: Left: Discretized PDDE example with lower bound for solutions with $\|u(0)\|_2 = 1$. Notice that one but not both of the plotted solutions for the discretized partial DDE have supremum above the given lower bound. Right: Linearization of laser example with lower bound for solutions with $\|u(0)\|_2 = 0.0015\|(E_x, E_y, N)^T\|_2$. The solution to the linearized system from the laser model departs significantly from the equilibrium before decaying asymptotically. Unless the truncated nonlinear part of the original laser model is guaranteed to stay small under departures which differ from the equilibrium by 0.38242 in norm, the applicability of the linear stability analysis to this equilibrium and these initial conditions is questionable.

4.6 Conclusion

Some practical, pointwise upper bounds on solutions to higher-order ODEs and single, constant delay DDEs have been demonstrated on a discretized partial DDE and a DDE model of a semiconductor laser with phase-conjugate feedback. A general lower bound was stated and used to concretely bound worst-case transient growth for both examples with a small computational effort. Effective techniques for localizing eigenvalues rather than computing them were used in an auxiliary fashion.

CHAPTER 5

SCATTERING RESONANCES AND QUANTUM CORRALS

5.1 Introduction

In 1993, IBM researchers Crommie, Lutz, and Eigler used a scanning tunneling microscope (STM) to place iron atoms into a circle on a copper surface [CLE93]. With the STM, they were able to create images not only of this "corral", but also of the quantum states of electrons trapped inside it.[1] Although admitting that failure to include the possibility of "transmission past the boundary atoms" [CLE93, p. 220], e.g. quantum tunneling, could be a source of error, they nevertheless successfully explained the qualitative behavior of observed quantum states using a two-dimensional, circular version of the first quantum model every physics student is introduced to: the particle in a box[2] [Gri05]. Within a few years, others had attempted to explain quantum corral experiments with more sophisticated models, including multiple scattering [CB96, FH03, HCLE94], elastic scattering with various potentials [HP96, RZ04], computations based on analogies with acoustics [BZH10], and a joint wave packet propagation approach where the corral is considered a circular wire [DTOBG08, GDTB09]. There has also been theoretical work on the distribution of resonances for quantum corrals, where the potential due to the corral atoms is modeled as a delta function on a circle [Gal16].

The success of the approaches cited above (multiple scattering, elastic scattering, etc.) is usually measured by how well the model used can reproduce

[1] `http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=4252` contains a gallery of high resolution images.

[2] This model is also called the infinite square well.

experimental observations of the *local density of states* (LDOS).[3] This quantity, which depends on both a position $\mathbf{x}$ and an energy E, relates to the likelihood of finding an electron with energy E at location $\mathbf{x}$. The actual quantity the STM measures, which is supposed to reflect the local density of states, is the differential conductance dI/dV [RZ04, Eq. (1)], i.e. the derivative of a measured current with respect to an applied voltage. If a certain energy parameter is fixed, allowing the STM tip to move across a diameter of the corral produces curves of dI/dV for fixed energy and varying position, as in [HCLE94, Fig. 2].

Different authors have advocated for the use of multiple scattering theory or for elastic scattering theory, but it is not clear which more accurately reflects the underlying physics (see for example [RZ04, Fig. 1] and discussion in [HP96]). And those opting for elastic scattering vary in their approaches to modeling the potential. Some choices include modeling each corral atom with finite cylinders [HP96] or Gaussian bumps [RZ04], and taking care about placing their centers on adsorption sites for the copper metal lattice rather than modeling them as equally spaced around a circle or other curve [FH03, HP96, HCLE94, RZ04]. It is not clear how much these choices matter for the computation of resonances themselves, because these approaches have only been compared by how well they reproduce the observed LDOS.

We will outline a framework for a comparative analysis between models. The key goal is to be able to concretely bound the difference between resonances predicted by corral models with different potentials and/or boundary conditions. This comparison will take place in the context of the discretizations of the partial differential equations appearing in the elastic scattering model, allowing

[3] A concise resource about the STM and what it measures can be found at `http://davisgroup.lassp.cornell.edu/STM_theory.html`.

us to formulate the computation of quantum states and their energies as a nonlinear eigenvalue problem, and giving us access to a powerful combination of linear algebra and complex analysis tools.

A brief outline is as follows. In Section 5.2 we will introduce models which permit quantum tunneling, the definition of resonance, the Dirichlet-to-Neumann map, and some facts about Schur complements that we will need later. In Section 5.3, we will describe the details of the discrete formulations and resulting nonlinear eigenvalue problems to be compared. Section 5.4 contains two types of error analysis: one based on first-order perturbation theory and the other on localization theorems for matrix-valued functions (Chapter 2 and [BH13]). Various sources of error due to the discretizations will be dealt with through these means. In Section 5.5 we apply these tools to compare both the particle-in-a-box (5.1) model and a formulation using rational approximations to a model where an exact scattering boundary condition is used.

5.2 Background

Mathematically, the particle in a box is posed as the following partial differential equation with Dirichlet boundary conditions:

$$\begin{aligned} -\Delta\psi &= E\psi \qquad \text{on } B(0,R) \\ \psi &= 0 \qquad\quad \text{on } r = R \end{aligned} \tag{5.1}$$

where we have used a scaling to eliminate the usual constants in the Schrödinger equation and R is the radius of the corral. Separation of variables shows $\psi(r,\theta) = \sum_n a_n J_n(kr)e^{in\theta}$, where $k^2 = E$ and each J_n is a first-kind Bessel function. To respect the convention that the first kind Hankel function $H^{(1)}(kr)$

is outgoing, k is taken as the principal square root of E. Eigenstates and their corresponding energies are then extracted by computing the zeros of $J_n(kR)$ for each $n \in \mathbb{Z}$. While the payoff from this naïve analysis in [CLE93] is quite good, one wonders *how much* the failure to account for quantum tunneling explains observed disagreement between calculation and experiment.

Since quantum tunneling through the corral is possible, the waves observed within the corral should be characterized as resonant (long-lived but transient) states rather than eigenstates [DZ16]. The resonances of a quantum system are defined to be those complex numbers $E = k^2$ such that

$$(-\Delta + V - E)\psi = 0 \qquad \text{(Time-independent Schrödinger equation)} \quad (5.2a)$$

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial}{\partial r} - ik\right)\psi = 0 \qquad \text{(Sommerfeld radiation condition)} \quad (5.2b)$$

has a nonzero solution ψ, where V is the potential due to the corral. In this case, the solution ψ is known as a resonant state, and the imaginary part of E is inversely proportional to the lifetime of the state. The particle-in-a-box model uses $V(r) = 0$ for $r < R$ and $V(r) = \infty$, simplifying (5.2) to an eigenvalue problem for a self-adjoint operator.

The potential induced by the quantum corral is necessarily of limited strength, as is the extent of its influence. The simplest model potential function that reflects these qualities is

$$V(r) = \begin{cases} 0, & r < R_1 \\ V_0, & R_1 < r < R_2 \\ 0, & r > R_2 \end{cases}, \qquad (5.3)$$

where V_0 is a positive constant and $[R_1, R_2]$ is the annulus on which the potential is supported (see Figure 5.2). Other options for V include those used in [HP96]

or [RZ04]. For simplicity, we will use (5.3) through Section 5.4.

It turns out that the rather unwieldy condition (5.2b) can be simplified under the assumption that the potential has compact support in $B(0,R)$. In particular, it can be rephrased as a boundary condition at $r = R$ by using the Dirichlet-to-Neumann (DtN) map, a pseudodifferential operator mapping solution values on the boundary to derivatives on the boundary [Giv99]. On the circular boundary $r = R$ (see Appendix C), the action of the DtN map $f(k,R)$ corresponding to (5.2) can be written in terms of Fourier expansions as

$$\sum_n c_n e^{in\theta} \xmapsto{f(k,R)} \sum_n f_n(k,R) c_n e^{in\theta}, \qquad f_n(k,R) = k\frac{\left(H_n^{(1)}\right)'(kR)}{H_n^{(1)}(kR)}, \tag{5.4}$$

where $H_n^{(1)}$ is the outgoing Hankel function of order n [AW05, §11.4].

This is a good time to simplify (5.2a) based on axisymmetry of (5.3) as well as motivate how the DtN map comes into play. First of all, using separation of variables on (5.2a), it can be shown that each component of the Fourier series $\psi(r,\theta) = \sum_n \psi_n(r)e^{in\theta}$ of the resonant state satisfies

$$-\psi_n''(r) - \frac{1}{r}\psi_n'(r) + \left(\frac{n^2}{r^2} + V(r) - E\right)\psi_n(r) = 0, \quad n \in \mathbb{Z}, \tag{5.5}$$

which reduces to Bessel's equation

$$r^2\psi_n''(r) + r\psi_n'(r) + \left((kr)^2 - n^2\right)\psi_n(r) = 0 \qquad \left(E = k^2\right) \tag{5.6}$$

outside of $B(0,R)$. Since the outgoing and incoming Hankel functions $H_n^{(1)}(kr)$ and $H_n^{(2)}(kr)$ (resp.) are a basis for the solution space of (5.6) [AW05, §11.4], and the radiation boundary condition (5.2b) picks out only the outgoing $H_n^{(1)}(kr)$ term, each $\psi_n(r)$ looks like a multiple of $H_n^{(1)}(kr)$ away from $B(0,R)$. As a consequence, satisfying (5.2b) is equivalent to the condition $\psi_n'(R) = f_n(k,R)\psi_n(R)\ \forall n$, or $\psi_r(R,\theta) = f(k,R)\psi(R,\theta)\ \forall\theta$, which is where the Dirichlet-to-Neumann map gets its name.

Having split (5.2) with (5.3) into a sequence of 1D problems, for now we will focus on one at a time, the n-th being

$$\begin{aligned} \left(-\frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} + \frac{n^2}{r^2} + V(r) - E\right)\psi_n(r) = 0 \qquad & \text{on } B(0,R) \\ \psi_n'(R) = k\frac{\left(H_n^{(1)}\right)'(kR)}{H_n^{(1)}(kR)}\psi_n(R). & \end{aligned} \tag{5.7}$$

To be clear, the resonances of (5.2) are precisely those complex numbers E such that that (5.7) has a nonzero solution for at least one index n.

In the next section, we will discretize (5.7) to approximate the operator in (5.7) by a matrix-valued function $T_n^{(\mathrm{DtN})}(k)$ (a different one for each index n). Recall that we define an eigenvalue of a matrix-valued function $T : \Omega \subset \mathbb{C} \to \mathbb{C}^{M\times M}$ as a complex number λ such that $T(\lambda)$ is a singular matrix. Therefore, the resonances of (5.7) will be approximated by $E = k^2$ where k is an eigenvalue of some $T_n^{(\mathrm{DtN})}$. It will turn out that $T_n^{(\mathrm{DtN})}$ is highly nonlinear and the problem of computing its eigenvalues cannot be attacked directly (although $T_n^{(\mathrm{DtN})}$ itself remains invaluable for error analysis in Section 5.4). Instead, we will obtain resonance approximations from linear matrix-valued functions $z \mapsto A - zB$ that have Schur complements approximating $T_n^{(\mathrm{DtN})}$. We will also use Schur complements to relate problems posed on different domains. We now define the Schur complement and some related vocabulary.

Consider the block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{5.8}$$

If A_{22} is nonsingular, then a block Gaussian elimination gives

$$\begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ A_{21} & A_{22}. \end{bmatrix} \tag{5.9}$$

Then it is customary to say that the matrix $A_{11} - A_{12}A_{22}^{-1}A_{21}$ is the Schur complement of A_{22} in A [Zha05]. Informally, we may also say that $A_{11} - A_{12}A_{22}^{-1}A_{21}$ is the Schur complement of A onto its leading block. Alternatively, if A is operating on a vector of variables $[x_1^T, x_2^T]^T$, then $A_{11} - A_{12}A_{22}^{-1}A_{21}$ acts on x_1 alone. Thus we also sometimes say that we have Schur complemented away the variables in x_2.

If $T : \Omega \to \mathbb{C}^{n\times n}$ is a matrix-valued function, we can also define a Schur complement $S(z) = T_{11}(z) - T_{12}(z)T_{22}(z)^{-1}T_{21}(z)$ by an appropriate partition. Notice that S is now a matrix-valued function as well. Generically speaking, T and S will have the same eigenvalues.

Lemma 5.1. *Let $T : \mathbb{C} \to \mathbb{C}^{M\times M}$ be a matrix-valued function, let I_1, I_2 be complementary subsets of $\{1, 2, ..., M\}$ and define $T_{i,j}(z) := T(z)_{I_i,I_j}$, $i, j = 1, 2$. If λ is not an eigenvalue of $T_{2,2}$, then $T(\lambda)$ is singular if and only if $T_{1,1}(\lambda) - T_{1,2}(\lambda)T_{2,2}(\lambda)^{-1}T_{2,1}(\lambda)$ is singular.*

Proof. For simplicity, suppose that I_1 is the index vector $(1, 2, ..., p)$ and $I_2 = (p + 1, p + 2, ..., M)$. Other cases can be reduced to this one by conjugating T by a permutation matrix.

Let λ be an eigenvalue of T. Then there is some nonzero $x = [x_1^T, x_2^T]^T$ such that

$$\begin{bmatrix} T_{11}(\lambda) & T_{12}(\lambda) \\ T_{21}(\lambda) & T_{22}(\lambda) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5.10}$$

If λ is not an eigenvalue of T_{22}, the above system is equivalent to

$$\begin{bmatrix} T_{11}(\lambda) - T_{12}(\lambda)T_{22}(\lambda)^{-1}T_{21}(\lambda) & 0 \\ T_{21}(\lambda) & T_{22}(\lambda) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5.11}$$

Since $T_{22}(\lambda)$ is nonsingular, x_1 must be nonzero. Thus, $T(\lambda)$ singular implies $T_{1,1}(\lambda) - T_{1,2}(\lambda)T_{2,2}(\lambda)^{-1}T_{2,1}(\lambda)$ singular. The converse follows from setting $x_2 = -T_{22}(\lambda)^{-1}T_{21}(\lambda)$. □

5.3 Discretization

There are a several options for discretizing (5.7). A popular choice is finite elements, used in [HP96]. See also [Lin02] for a variational discretization method used with complex scaling [Moi98] and an interesting choice of basis. Here we use spectral collocation [Tre00], which is known to exhibit fast convergence as the number of discretization points is increased. Applying spectral collocation to (5.7) amounts to approximating the solution $\psi_n(r)$ by a polynomial $p_n(r)$ and enforcing that p_n satisfies (5.7) at all points in a certain mesh of $[0, R]$.

The first step in setting up a spectral collocation discretization is to choose the mesh. To capture the discontinuity in (5.3), we will mesh $[0, R_1]$ and $[R_1, R_2]$ separately, the first with a half Chebyshev mesh $\hat{r}_1$ to avoid redundancy at the origin [Tre00, Ch. 11], and the second with the usual Chebyshev extreme points $\hat{r}_2$ (see Figure 5.2 for a cartoon of the mesh on $[0, R_1]$ and $[R_1, R_2]$). Note that this means $\hat{r}_1$ and $\hat{r}_2$ both include the point R_1. Continuity and differentiability of $p_n(r)$ will be enforced at the interface point R_1, and the DtN map boundary condition from (5.7) will be enforced at $R = R_2$.

Let D_1 be the Chebyshev differentiation matrix mapping values of $p_n(r)$ to values of $p_n'(r)$ on $\hat{r}_1$ and define D_2 similarly (see [Tre00] for details and code). Let $\hat{r}$ be the concatenation of $\hat{r}_1$ and $\hat{r}_2$. For notational convenience, if $\hat{x} = [x_1, x_2, ..., x_M]^T$ is a vector of points and f is a function, then we let

$f(\hat{x})$ denote the vector $[f(x_1), f(x_2), ..., f(x_M)]^T$. With this convention, we write $D_j p_n(\hat{r}_j) = p_n'(\hat{r}_j)$, $j = 1, 2$. As a further notational convenience let $\text{diag}(\hat{x})$ denote the diagonal matrix with diagonal entries from the vector $\hat{x}$, and let $1/\hat{x}$ and $\hat{x}^s$ be understood as the vectors $[1/x_1, 1/x_2, ..., 1/x_M]^T$ and $[x_1^s, x_2^s, ..., x_M^s]^T$, respectively (akin to MATLAB syntax). Then, ignoring boundary conditions at R_1 and R_2 for the moment, the discretization of (5.7) is $\left(A_{n,j} - k^2 I\right) p_n(\hat{r}_j) = 0$, $j = 1, 2$, where $A_{n,j} = -D_j^2 - \text{diag}(1/\hat{r}_j) D_j + \text{diag}(n^2/\hat{r}_j^2) + \text{diag}(V(\hat{r}_j))$. Enforcing interface and boundary conditions is done by replacing the rows corresponding to R_1 and R_2, yielding the discretization

$$\underbrace{(A_n^{(\text{DtN})} - k^2 B^{(\text{DtN})} + C_n^{(\text{DtN})}(k))}_{T_n^{(\text{DtN})}(k)} p_n(\hat{r}) = 0, \tag{5.12}$$

where

$$A_n^{(\text{DtN})} = \left[\begin{array}{c|c} A_{n,1}(1 : \text{end} - 1, :) & \\ -D_1(\text{end}, :) & D_2(1, :) \\ \hline 0 \quad \dots \quad 0 \quad 1 & -1 \quad 0 \quad \dots 0 \\ & A_{n,2}(2 : \text{end} - 1, :) \\ & -D_2(\text{end}, :) \end{array}\right], \tag{5.13}$$

$$B^{(\text{DtN})} = \text{diag}([1 \ \dots \ 1 \ 0 \,|\, 0 \ 1 \ \dots \ 1 \ 0]),$$

$$C_n^{(\text{DtN})}(k) = \text{diag}([0 \ \dots \ 0 \,|\, 0 \ \dots \ 0 \ f_n(k, R)]).$$

The expression $A_{n,1}(1 : \text{end} - 1, :)$ is MATLAB notation meaning all columns and all but the last row of $A_{n,1}$.

It is difficult to find eigenvalues of $T_n^{(\text{DtN})}$ due to the presence of the highly nonlinear term $C_n^{(\text{DtN})}(k)$. However, we may approximate (5.12) by something more tractable. To this end, in Section 5.5.2 we replace $C_n^{(\text{DtN})}(k)$ with a rational approximation $C_n^{(\text{rat})}(k) := -A_{n,12}\left(A_{n,22} - k^2 B_{22}\right)^{-1} A_{n,21}$. Then, the eigenvalues of

$T_n^{(\mathrm{rat})}(k) := A_n^{(\mathrm{DtN})} - k^2 B^{(\mathrm{DtN})} + C_n^{(\mathrm{rat})}(k)$ are all eigenvalues of

$$\left(T_{\mathrm{full}}^{(\mathrm{rat})}\right)_n(k) = \underbrace{\begin{bmatrix} A_n^{(\mathrm{DtN})} & A_{n,12} \\ A_{n,21} & A_{n,22} \end{bmatrix}}_{\left(A_{\mathrm{full}}^{(\mathrm{rat})}\right)_n} - k^2 \underbrace{\begin{bmatrix} B^{(\mathrm{DtN})} & \\ & B_{22} \end{bmatrix}}_{B_{\mathrm{full}}^{(\mathrm{rat})}} \tag{5.14}$$

(by Lemma 5.1), which are easy to compute. And wherever $C_n^{(\mathrm{rat})}(k)$ approximates $C_n^{(\mathrm{DtN})}(k)$ closely enough, we will be able obtain approximate eigenvalues of $T_n^{(\mathrm{DtN})}$ through use of the pseudospectral inclusion result Theorem 2.5. This will be further discussed in Section 5.4.

A very popularly known rational approximation to the DtN map comes from using perfectly matched layers (PMLs) [Kim09]. While there are physical ways of describing PMLs and how to select good parameters, the mathematical description is a complex-valued coordinate change that bends $[R, \infty)$ into the upper half of the complex plane. This causes the waves $H_n^{(1)}$ and $H_n^{(2)}$ to be attenuated and amplified, respectively, as distance from the potential support increases. Recalling that each ψ_n must be a multiple of $H_n^{(1)}$ for $r > R$, a Dirichlet boundary condition applied sufficiently far from the potential support serves as a reasonable way to select for $H_n^{(1)}$ after sufficient attenuation has occurred. After discretizing and Schur complementing away the variables associated to the PML region (outside the potential support and up to the radius where a Dirichlet boundary was enforced), one obtains $T_n^{(\mathrm{PML})}(k) = A_n^{(\mathrm{DtN})} - k^2 B^{(\mathrm{DtN})} + C_n^{(\mathrm{PML})}(k)$. It tends to be that $C_n^{(\mathrm{PML})}(k)$ approximates $C_n^{(\mathrm{DtN})}(k)$ best for moderately-sized k, since accuracy is lost both near the origin due to the square root branch cut and far from the origin due to numerical reflection artifacts. To enlarge the region of accuracy, the number of discretization points in the PML region can be increased, but accuracy will always be best for moderate k.

In contrast, we can choose other rational approximations to be most accurate wherever we want. The following proposition shows how to obtain such a rational approximation for a range of energies k^2.

Proposition 5.1. *Let $f(z) = f_n(\sqrt{z}, R)$ using the principal branch of the square root and let N be a positive integer. Let Ω be a chosen domain of interest with smooth, positively oriented boundary parametrized by $\varphi : [0,1] \to \partial\Omega$, and suppose $f(z)$ is analytic in $\bar{\Omega}$. Define*

$$\begin{aligned} z_j &= \varphi(j/N), \\ w_j &= f(z_j)\varphi'(j/N)\frac{2/N}{4\pi i}, \quad j = 1, ..., N-1, \\ w_j &= f(z_j)\varphi'(j/N)\frac{1/N}{4\pi i}, \quad j = 0, N \end{aligned} \tag{5.15}$$

and let $\hat{z}$ and $\hat{w}$ be column vectors of the poles z_j and nodes w_j as defined above. Define

$$\left(A^{(rat)}_{full}\right)_n = \left[\begin{array}{ccc|ccc} & & & 0 & \dots & 0 \\ & A_n^{(DtN)} & & \vdots & & \vdots \\ & & & 0 & \dots & 0 \\ & & & & \hat{w}^T & \\ \hline 0 & \dots & 0 & \ddots & & \\ \vdots & & \vdots \; -\mathbf{1} & & \mathrm{diag}(\hat{z}) & \\ 0 & \dots & 0 & & & \ddots \end{array}\right], \qquad B^{(rat)}_{full} = \left[\begin{array}{c|c} B^{(DtN)} & \\ \hline & I \end{array}\right] \tag{5.16}$$

where $\mathbf{1}$ represents the vector of all ones, and define $T_n^{(rat)}(k)$ as the Schur complement of $\left(A^{(rat)}_{full}\right)_n - k^2 B^{(rat)}_{full}$ onto the leading block. Then for any k in the right half-plane with $k^2 \in \Omega$, $T_n^{(rat)}(k) \to T_n^{(DtN)}(k)$ as $N \to \infty$.

Proof. The Cauchy integral formula states that under the conditions of the theorem

$$f(k^2) = \frac{1}{2\pi i} \oint_{\partial\Omega} \frac{f(z)}{z - k^2}\, dz \tag{5.17}$$

for any $k^2 \in \Omega$. Since the principal branch of the square root is being used, $\sqrt{k^2} = k$ if and only if k is in the left half-plane. The result follows by approximating the integral using the trapezoid rule. □

5.4 Error analysis

We have taken several steps getting from the definition of the resonance in the continuous, two-dimensional setting (5.2) to the kind of problem we can solve numerically, and most introduced a new source of error. In this section we continue to use the axisymmetric potential (5.3), so that moving from (5.2) to the full sequence of 1D problems (5.7) preserves eigenvalues exactly. But the following steps do not:

1. Passing from (5.7) to the discretized version (5.12).
2. Solving an approximating, linear eigenvalue problem (5.14) instead of the difficult nonlinear eigenvalue problem (5.12) for a genuinely nonlinear matrix-valued function.
3. Neglecting (5.12) for infinitely many n.

5.4.1 Error due to discretization

To address the first source of error, recall that the spectral collocation discretization amounts to replacing ψ_n with a polynomial approximant $p_n(r)$ in (5.7), and subsequently computing k^2 and the values of some nonzero $p_n(r)$ on a mesh. Then a sum such as $\sum_{n=N_1}^{N_2} p_n(r)e^{in\theta}$ (usually $N_1 = -N_2$ or $-N_2 + 1$) approximates

the solution to (5.2). In any case, suppose $\hat{\psi}, \hat{k}$ is some approximate solution pair for (5.2), written here for simplicity of notation as

$$\begin{cases} \left(H - k^2\right)\psi = 0 & \text{on } B(0,R) \\ \left(\frac{\partial}{\partial r} - f(k,R)\right)\psi = 0 & \text{on } \partial B(0,R) \end{cases} \tag{5.18}$$

where $H = -\Delta + V$ and the Sommerfeld boundary condition has been replaced with the DtN map boundary condition on $r = R$. For $\hat{\psi}, \hat{k}$ to be an approximate solution, there must exist $\delta\psi$ with small $L^2(B(0,R))$ norm and small δk such that $\psi = \hat{\psi} + \delta\psi$ and $k = \hat{k} + \delta k$. From another point of view, the approximate pair must have a small residual, i.e., $(H - \hat{k}^2)\hat{\psi}$ should be small on $B(0,R)$ and $(\partial/\partial r - f(\hat{k},R))\hat{\psi}$ should be small on $r = R$. Then a residual-based error estimate for the eigenvalues computed by discretizing (5.2) is contained in the next proposition, for which the following lemma is essential.

Lemma 5.2. *If $u, v \in H^1(B(0,R))$, then $\int_{\partial B(0,R)} u f(k,R) v = \int_{\partial B(0,R)} v f(k,R) u$.*

Proof. See Appendix C. □

Proposition 5.2. *Suppose*

$$\begin{cases} \left(H - \hat{k}^2\right)\hat{\psi} = \eta & \text{on } B(0,R) \\ \left(\frac{\partial}{\partial r} - f(\hat{k},R)\right)\hat{\psi} = \alpha & \text{on } \partial B(0,R). \end{cases} \tag{5.19}$$

Then to first order,

$$\delta k \approx \frac{\int_{B(0,R)} \hat{\psi}\eta + \int_{\partial B(0,R)} \hat{\psi}\alpha}{2\hat{k}\int_{B(0,R)} \hat{\psi}^2 + \int_{\partial B(0,R)} \hat{\psi} f_k(\hat{k},R)\hat{\psi}}. \tag{5.20}$$

Remark 5.1. Since in our discretization the values of $\hat{\psi}$ are known on a Chebyshev-based mesh in the radial direction, integration with respect to r can be performed by applying the Chebfun [DHT14] `sum()` function.

Proof. Deleting second-order terms from $\left(H - (\hat{k} + \delta k)^2\right)\left(\hat{\psi} + \delta\psi\right) = 0$ and $\left(\partial/\partial r - f(\hat{k} + \delta k, R)\right)\left(\hat{\psi} + \delta\psi\right) = 0$ shows that to first order the δk, $\delta\psi$ pair satisfies

$$\begin{cases} \left(H - \hat{k}^2\right)\delta\psi - 2\hat{k}\delta k\hat{\psi} = -\eta & \text{on } B(0,R) \\ \left(\frac{\partial}{\partial r} - f(\hat{k}, R)\right)\delta\psi - f_k(\hat{k}, R)\hat{\psi}\delta k = -\alpha & \text{on } \partial B(0,R). \end{cases} \tag{5.21}$$

Applying Green's identity [Eva98, p. 628, Theorem 3(iii)] (essentially an integration by parts) gives

$$\int_{B(0,R)} \hat{\psi}\Delta(\delta\psi) = \int_{B(0,R)} (\delta\psi)\Delta\hat{\psi} + \int_{\partial B(0,R)} \left(\hat{\psi}(\delta\psi)_r - (\delta\psi)\hat{\psi}_r\right), \tag{5.22}$$

and

$$\int_{B(0,R)} \hat{\psi}(H - \hat{k}^2)\delta\psi = \int_{\partial B(0,R)} \left((\delta\psi)\hat{\psi}_r - \hat{\psi}(\delta\psi)_r\right). \tag{5.23}$$

Using (5.21) in the previous equation we obtain

$$\int_{B(0,R)} \hat{\psi}(2\hat{k}\delta k - \eta) = \int_{\partial B(0,R)} \left[\delta\psi f(\hat{k}, R)\hat{\psi} - \hat{\psi}\left(f(\hat{k}, R)\delta\psi + f_k(\hat{k}, R)\hat{\psi}\delta k - \alpha\right)\right]. \tag{5.24}$$

By the lemma, this reduces to

$$\int_{B(0,R)} \hat{\psi}(2\hat{k}\delta k\hat{\psi} - \eta) = \int_{\partial B(0,R)} \hat{\psi}\alpha - \int_{\partial B(0,R)} \hat{\psi} f_k(\hat{k}, R)\hat{\psi}\delta k, \tag{5.25}$$

and solving for δk gives the result. □

Corollary 5.1. *If $\hat{\psi}(r,\theta) = \sum_{n=N_1}^{N_2} \hat{\psi}_n(r)e^{in\theta}$ and we express all other functions in terms of their Fourier series as well, then by Parseval's theorem*

$$\delta k = \frac{\sum_{n=N_1}^{N_2} \int_0^R \hat{\psi}_n(r)\eta_{-n}(r)\, r\, dr + 2\pi R \sum_{n=N_1}^{N_2} \hat{\psi}_n(R)\alpha_{-n}}{2\hat{k}\sum_{n=N_1}^{N_2} \int_0^R \hat{\psi}_n(r)\hat{\psi}_{-n}(r)\, r\, dr + 2\pi R \sum_{n=N_1}^{N_2} \hat{\psi}_n(R)(f_{-n})_k(\hat{k}, R)\hat{\psi}_{-n}(R)}. \tag{5.26}$$

5.4.2 Error due to rational approximation

To address the second source of error, we can also do a first-order analysis, but this time entirely in the discretized framework. In the following proposition, we

think of T as being the exact problem and $\hat{T}$ as an approximating problem that is easier to solve.

Proposition 5.3. *Let $T(z)$ and $\hat{T}(z)$ be matrix-valued functions satisfying $T(z) = \hat{T}(z) + E(z)$. Suppose we have an eigentriple $\hat{w}, \hat{v}, \hat{\lambda}$ for $\hat{T}$, i.e. $\hat{w}^* \hat{T}(\hat{\lambda}) = \hat{T}(\hat{\lambda})\hat{v} = 0$, and suppose $w = \hat{w} + \delta w$, $v = \hat{v} + \delta v$, $\lambda = \hat{\lambda} + \delta\lambda$ is an eigentriple for T. If δw, δv, $\delta\lambda$, and $E(\hat{\lambda})$ are all small, and if $\hat{T}$ and T are both analytic at $\hat{\lambda}$, then to first order*

$$\delta\lambda = -\frac{\hat{w}^* E(\hat{\lambda})\hat{v}}{\hat{w}^* T'(\hat{\lambda})\hat{v}}. \tag{5.27}$$

Proof. By assumption, $(\hat{w}^* + \delta w^*)\left(\hat{T}(\lambda) + E(\lambda)\right)(\hat{v} + \delta v) = 0$. Also, $\hat{T}(\lambda) \approx \hat{T}(\hat{\lambda}) + \hat{T}'(\hat{\lambda})\delta\lambda$ and $E(\lambda) \approx E(\hat{\lambda}) + E'(\hat{\lambda})\delta\lambda$. Putting them together we obtain

$$\begin{aligned} 0 &= (\hat{w}^* + \delta w^*)\left(\hat{T}(\hat{\lambda}) + E(\hat{\lambda}) + T'(\hat{\lambda})\delta\lambda\right)(\hat{v} + \delta v) \\ &\approx \hat{w}^* E(\hat{\lambda})\hat{v} + \hat{w}^* T'(\hat{\lambda})\hat{v}\delta\lambda \end{aligned} \tag{5.28}$$

after dropping higher-order terms and terms that are equal to zero. Solving for $\delta\lambda$ gives the result. □

In some cases it is possible to get concrete error bounds through Theorem 2.5 as well, as we will demonstrate in Section 5.5. Recall that the gist of the pseudospectral inclusion result in Theorem 2.5 is as follows. Suppose that $\hat{\lambda}$ is the only eigenvalue of $\hat{T}$ in a bounded, connected component $\mathcal{U}$ of the ε-pseudospectrum of T, where $\varepsilon > 0$ is some small parameter. Then if $\|T - \hat{T}\| < \varepsilon$ on $\bar{\mathcal{U}}$ and $T, \hat{T}$ are both analytic on $\mathcal{U}$, then there is exactly one eigenvalue of T in $\mathcal{U}$. Similarly, if there are n eigenvalues of $\hat{T}$ in $\mathcal{U}$, then there must be exactly n eigenvalues of T in $\mathcal{U}$ as well.

5.4.3 Truncation of the sequence of 1D problems

Now we deal with the third and last source of error, due to truncating the infinite sequence of problems (5.12) for $n \in \mathbb{Z}$. The concern is that even if we specify a bounded domain Ω of interest and verify that any eigenvalues we have computed there are accurate, we may have missed some due to dropping the problems associated to large $|n|$. In this section, we prove using the Gershgorin generalization Theorem 2.1 that we needn't worry.

It is clear from (5.13) and the definition of $A_n^{(\mathrm{DtN})}$ that the matrix-valued functions $T_n^{(\mathrm{DtN})}$, $n \in \mathbb{Z}$, differ from each other only in the $n^2/\hat{r}^2$ term in their respective $A_n^{(\mathrm{DtN})}$'s, and in their $C_n^{(\mathrm{DtN})}$ terms. To make this more explicit, let us express $A_n^{(\mathrm{DtN})}$ as $A + D_n$, where D_n equals $\mathrm{diag}(n^2/\hat{r}^2)$ with the two rows corresponding to $r = R_1$ and the one row corresponding to $r = R_2$ set to zero. Then $T_n^{(\mathrm{DtN})}(k) = A + D_n - k^2 B^{(\mathrm{DtN})} + C_n^{(\mathrm{DtN})}(k)$. Now, let I_2 be an index vector corresponding to the two rows corresponding to $r = R_1$, where D_n and $B^{(\mathrm{DtN})}$ are zero (see (5.13)), and let I_1 be the index vector for the remaining rows. Taking advantage of MATLAB notation once again, denote the submatrix consisting of the I_i rows and I_j columns of a matrix M by $M(I_i, I_j)$. Now, we can Schur complement away the I_2 variables in $T_n^{(\mathrm{DtN})}(k)$ to obtain

$$S_n(k) = A(I_1, I_1) + \begin{bmatrix} \mathrm{diag}(n^2/\hat{r}(I_1)^2) - k^2 I & \\ & f_n(k, R) \end{bmatrix} - A(I_1, I_2)A(I_2, I_2)^{-1}A(I_2, I_1). \tag{5.29}$$

Then the following theorem shows that for a fixed mesh $\hat{r}$ of $[0, R_2]$, we only need solve the nonlinear eigenvalue problem (5.12) for finitely many n.

Theorem 5.1. *Suppose that Ω is a bounded region in the complex plane, and define the matrix-valued functions $T_n^{(DtN)}$ and S_n as above. If $A(I_2, I_2)$ is nonsingular, then there*

exists N such that $T_n^{(DtN)}$ has no eigenvalues in Ω if $|n| \geq N$.

Proof. Let k be fixed. Recall that the outgoing Hankel function $H_n^{(1)}$ satisfies $H_n^{(1)}(z) = J_n(z) + iY_n(z)$ and $\left(H_n^{(1)}\right)'(z) = \left(H_{n-1}^{(1)}(z) - H_{n+1}^{(1)}(z)\right)/2$ [AW05, Ch. 11]. By Lemma C.2 in Appendix C,

$$f_n(k,R) \to -\frac{|n|}{R} + \frac{1}{|n|-1}\frac{k^2R}{4} \tag{5.30}$$

as $|n| \to \infty$. Thus, given $\varepsilon > 0$ and a fixed value of k, there exists $N_{k,\varepsilon}$ such that $|-|n|/R - f_n(k,R)| < \varepsilon$ for all $|n| \geq N_{k,\varepsilon}$.

Now, by Lemma 5.1 and the assumption that $A(I_2, I_2)$ is nonsingular, $S_n(k)$ and $T_n^{(\mathrm{DtN})}$ have the same eigenvalues. By the Gershgorin generalization Theorem 2.1, all eigenvalues of S_n are contained in the union $\bigcup_j G_{n,j}$ of the Gershgorin regions $G_{n,j}$, $j = 1, 2, ..., M$, defined as

$$\begin{aligned} G_{n,j} &= \{k \in \mathbb{C} : |n^2/\hat{r}(I_1(j))^2 - k^2| \leq \rho_j\}, \quad j = 1, 2, ..., M-1 \\ G_{n,M} &= \{k \in \mathbb{C} : |f_n(k,R)| \leq \rho_M\} \end{aligned} \tag{5.31}$$

where S_n maps to $M \times M$ matrices and ρ_j is the sum of the absolute values of the entries in the j-th row of $A(I_1, I_1) - A(I_1, I_2)A(I_2, I_2)^{-1}A(I_2, I_1)$.

The first $M-1$ Gershgorin regions have finite set diameter and contain points $\pm n/\hat{r}(I_1(j))$ that diverge, and hence $G_{n,j} \cap \Omega = \varnothing$ for $|n|$ large enough. To prove the same is true for $G_{n,M}$, we argue by contradiction. Suppose for infinitely many $n \in \mathbb{Z}$ there is some $k_n \in \Omega$ such that $|f_n(k_n, R)| \leq \rho_M$. By the Bolzano-Weierstrass theorem, there is a convergent subsequence of the k_n's and hence a limit k_0. This implies that $|f_n(k_0, R)| \leq \rho_M$ for arbitrarily large $n \in \mathbb{Z}$. However, that contradicts the fact that $|f_n(k_0, R)| \to \infty$ as $n \to \infty$ by (5.30). Therefore no such sequence of k_n's can exist, and hence there is some N such that $|f_n(k,R)| > \rho_M$ for all $|n| \geq N$

and all $k \in \Omega$. With that we have shown that for $|n| \geq N$, $G_{n,N} \cap \Omega = \varnothing$ and the proof is complete. □

5.5 Examples

5.5.1 The particle in a box

As previously discussed, the classical particle in a box model for the quantum corral with radius R is (5.1), and the resonance approximations according to this model are computed from zeros of Bessel functions J_n. To compute resonance approximations from the discretized version of the problem, we use separation of variables to split (5.1) according to (5.5) and apply the Dirichlet boundary condition to each ψ_n. We then discretize the continuous problem for index n, i.e.,

$$\begin{aligned} \left(-\frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} + \frac{n^2}{r^2} - k^2\right)\psi_n(r) = 0 \qquad \text{on } [0, R] \\ \psi_n(R) = 0 \end{aligned} \tag{5.32}$$

to obtain the matrix equation

$$\underbrace{\left(A_n^{(\mathrm{Dir})} - k^2 B^{(\mathrm{Dir})}\right)}_{T_n^{(\mathrm{Dir})}(k)} p_n(\hat{r}) = 0 \tag{5.33}$$

where

$$\begin{aligned} A_n^{(\mathrm{Dir})} &= \begin{bmatrix} A_{n,1}(1 : \mathrm{end} - 1, :) \\ 0 \quad \ldots \quad 0 \quad 1 \end{bmatrix}, \\ B^{(\mathrm{Dir})} &= \mathrm{diag}([1 \ \ldots \ 1 \ 0]) \end{aligned} \tag{5.34}$$

according to the procedure that led us to (5.12), (5.13). Then resonance energy estimates are the eigenvalues $E = k^2$ of the pencil $\left(A_n^{(\mathrm{Dir})}, B^{(\mathrm{Dir})}\right)$ [GL96]. To verify

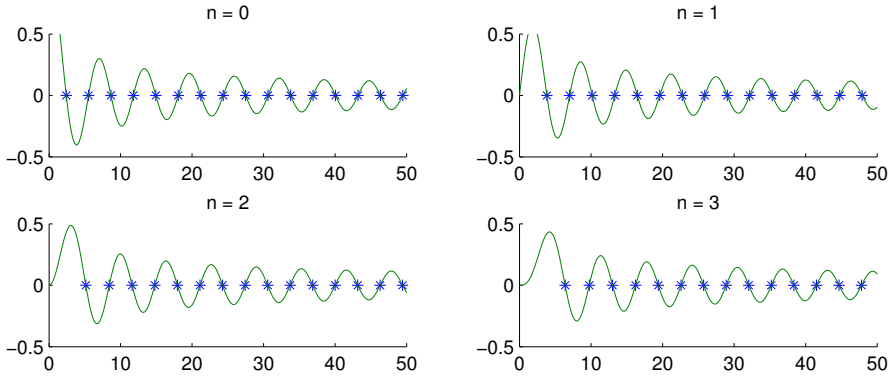

Figure 5.1: Eigenvalues of matrix-valued function $T_n^{(\mathrm{Dir})}(k)$ for index $n = 0, 1, 2, 3$, constructed using 40 and 10 points on $[0, R_1]$ and $[R_1, R]$, resp., for $R_1 = 0.95$, $R = 1$, with plot of corresponding Bessel function $J_n(kR)$. Resonance estimates are $E = k^2$ as defined by (5.32).

that the resonance estimates we compute through discretizing the particle in a box model are the same as those computed in [CLE93], consider Figure 5.1.

Recall that (5.32) is equivalent to putting

$$V(r) = \begin{cases} 0, & r < R, \\ \infty, & r > R \end{cases} \tag{5.35}$$

in (5.2). Now, suppose we would like to compare the output of the particle in a box model with some other model that permits quantum tunneling. As a simple case, take (5.7) with potential (5.3), where $R = 1$, $V_0 = 430$, $R_1 = R - w/2$ and $R_2 = R + w/2$, and $w = 0.1$. These parameters are chosen to make the potential a high wall concentrated around radius R in an attempt to be consistent with (5.35). In this section, $T_n^{(\mathrm{DtN})}$ will denote the discretization of (5.7) with these parameters. The potential for each of these models, and the underlying meshes of $[0, R_1]$, $[R_1, R]$, and $[R_1, R_2]$ are illustrated in Figure 5.2.

The question is, what is the connection between the resonances of the particle in a box model and the quantum tunneling model we have just defined?

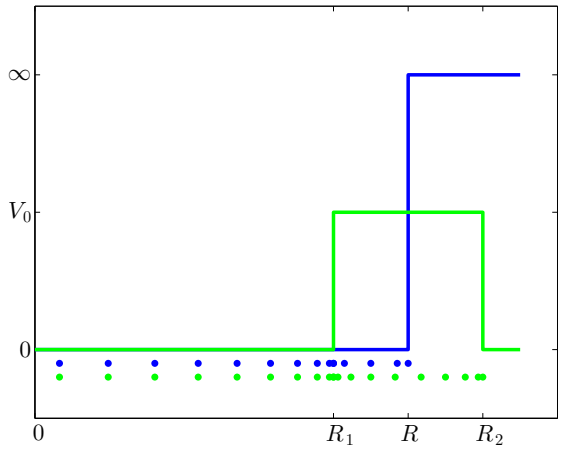

Figure 5.2: Illustration of two potentials with mutually consistent meshes.

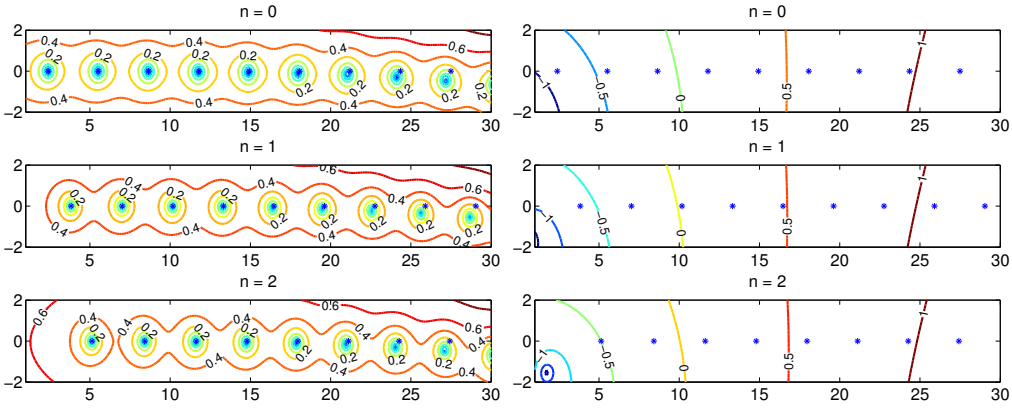

Figure 5.3: Pseudospectra and error norms used in comparing the particle in a box model to a quantum tunneling model.

In terms of matrix-valued functions, how do the eigenvalues of $T_n^{(\mathrm{Dir})}$ and $T_n^{(\mathrm{DtN})}$ relate? If we are to use Theorem 2.5 to answer, then we cannot use $T_n^{(\mathrm{Dir})}$ and $T_n^{(\mathrm{DtN})}$ directly because they are not the same size. Instead, we can Schur complement away the variables associated to $[R_1, R]$ and $[R_1, R_2]$, respectively, to obtain matrix-valued functions of the same size that act on vectors of values at the same mesh points.[4] Let us denote those Schur complements by $S_n^{(\mathrm{Dir})}$ and $S_n^{(\mathrm{DtN})}$.

Now let us examine Figure 5.3. On the left we have a 2-norm pseu-

[4] Caveat: to do this, we need to have meshed $[0, R_1]$, and $[R_1, R]$ separately, so that both $T^{(\mathrm{Dir})}$ and $T^{(\mathrm{DtN})}$ were created with the same mesh of $[0, R_1]$. Therefore (5.34) would be not quite correct, since we need C^1 boundary conditions at $r = R_1$ as well.

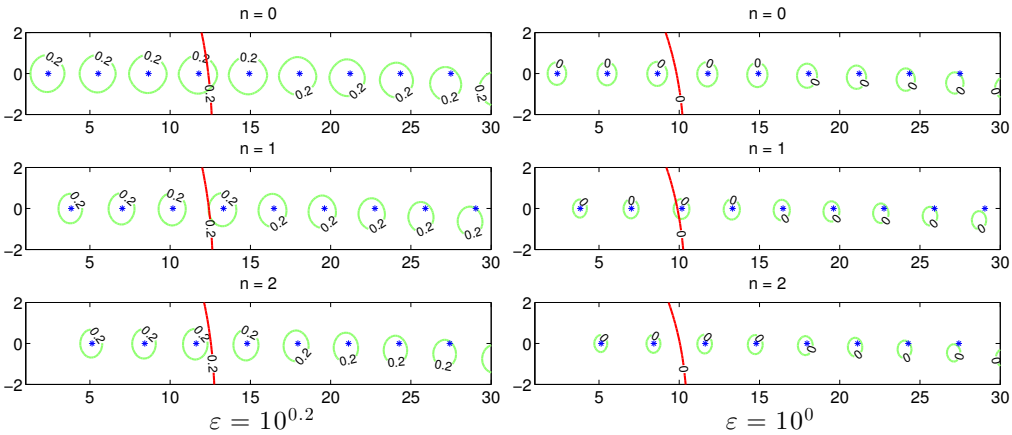

Figure 5.4: Plot used to deduce inclusion regions and counts for resonances of quantum tunneling model.

dospectral plot (Definition 1.5) of $S_n^{(\mathrm{DtN})}$ for $n = 0, 1, 2$, where contours of $\log_{10}\left(\sigma_{\min}(S_n^{(\mathrm{DtN})}(z))\right)$ are shown. Eigenvalues of $S_n^{(\mathrm{Dir})}$ are plotted as well. We can see that as we move from left to right the eigenvalues of $S_n^{(\mathrm{Dir})}$ become less accurate approximations for eigenvalues of $S_n^{(\mathrm{DtN})}$. On the right, we plot contours of $\log_{10} \|S_n^{(\mathrm{DtN})}(z) - S_n^{(\mathrm{Dir})}(z)\|_2$ for $n = 0, 1, 2$. By comparing the two plots, it appears that several components of the $10^{0.2}$-pseudospectrum of $S_n^{(\mathrm{DtN})}$ lie completely within the region where $S_n^{(\mathrm{DtN})}$ and $S_n^{(\mathrm{Dir})}$ differ by less than $10^{0.2}$ in 2-norm. The analogous fact appears to be true for the 10^0-pseudospectrum of $S_n^{(\mathrm{DtN})}$ and the region where $\|S_n^{(\mathrm{DtN})}(z) - S_n^{(\mathrm{Dir})}(z)\|_2 < 10^0$. Figure 5.4 (left) contains the $10^{0.2}$ contours of $\sigma_{\min}(S_n^{(\mathrm{DtN})}(z))$ (green) and $\|S_n^{(\mathrm{DtN})}(z) - S_n^{(\mathrm{Dir})}(z)\|_2$ (red) for $n = 0, 1, 2$. We deduce from Theorem 2.5 that any green component fully to the left of the red curve must contain exactly one resonance for the quantum tunneling model, i.e., one eigenvalue of $T_n^{(\mathrm{DtN})}$. Similar conclusions can be drawn about Figure 5.4 (right), where 10^0 contours are plotted.

We were able to localize a few of the smallest eigenvalues of $T_n^{(\mathrm{DtN})}$ for $n = 0, 1, 2$ this way, and the inclusion region components were reasonably small, so we have reason to believe that the first order correction derived in Propo-

n	$T^{(\mathrm{Dir})}$ eig $\hat{k}$	first order correction δk	$\hat{k}$ residual	$\hat{k}+\delta k$ residual
0	2.404826	$+0.000288 - 0.000911i$	$1.34e-04$	$1.01e-06$
0	5.520078	$-0.001711 - 0.004925i$	$6.72e-04$	$1.18e-05$
0	8.653728	$-0.009409 - 0.012977i$	$2.10e-03$	$6.92e-05$
0	11.791534	$-0.026079 - 0.026736i$	$5.29e-03$	$2.79e-04$
1	3.831706	$-0.000046 - 0.002246i$	$3.89e-04$	$4.28e-06$
1	7.015587	$-0.004338 - 0.008094i$	$1.49e-03$	$3.55e-05$
1	10.173468	$-0.015972 - 0.018662i$	$4.07e-03$	$1.70e-04$
1	13.323692	$-0.038196 - 0.036196i$	$9.55e-03$	$6.03e-04$
2	5.135622	$-0.000795 - 0.003902i$	$7.24e-04$	$1.05e-05$
2	8.417244	$-0.007913 - 0.011780i$	$2.50e-03$	$7.57e-05$
2	11.619841	$-0.024014 - 0.025243i$	$6.44e-03$	$3.25e-04$
2	14.795952	$-0.052291 - 0.047245i$	$1.46e-02$	$1.07e-03$
3	6.380162	$-0.001966 - 0.005857i$	$1.14e-03$	$2.05e-05$
3	9.761023	$-0.012441 - 0.016013i$	$3.78e-03$	$1.38e-04$
3	13.015201	$-0.033531 - 0.032818i$	$9.41e-03$	$5.53e-04$
3	16.223466	$-0.068326 - 0.060106i$	$2.11e-02$	$1.73e-03$

Table 5.1: Some eigenvalues of $T_n^{(\mathrm{Dir})}$ for various n and their first order corrections according to Proposition 5.3. The residual at k is defined to be $\sigma_{\min}\left(T^{(\mathrm{DtN})}(k)\right)$.

sition 5.3 may be reasonably small as well. Suppose that $\hat{k}$, $\hat{v}$ and $\hat{w}$ satisfy $S_n^{(\mathrm{Dir})}(\hat{k})\hat{v} = \hat{w}^* S_n^{(\mathrm{Dir})}(\hat{k}) = 0$. Then the first-order sensitivity analysis from Proposition 5.3 suggests that $\hat{k}+\delta k$ is approximately an eigenvalue of the matrix-valued function $S_n^{(\mathrm{DtN})}$ (and hence of $T_n^{(\mathrm{DtN})}$ by Lemma 5.1), where

$$\delta k = -\frac{\hat{w}^*\left(S_n^{(\mathrm{DtN})}(\hat{k}) - S_n^{(\mathrm{Dir})}(\hat{k})\right)\hat{v}}{\hat{w}^*\left(S_n^{(\mathrm{DtN})}\right)'(\hat{k})\hat{v}}. \tag{5.36}$$

See Table 5.1 for some numerical values of δk. In the particular case of eigenvalue $\hat{k} \approx 2.404826$ of $T_0^{(\mathrm{Dir})}$, applying five steps of bordered Newton [Gov00, Ch. 3] on $T_n^{(\mathrm{DtN})}$ using $\hat{k}$ as an initial guess we find that $k = 2.405119613525277 - 0.000906699509689i$ is very close to an eigenvalue of $T_n^{(\mathrm{DtN})}$ (residual $\approx 6.91 \times 10^{-13}$), and $k - \hat{k} \approx -0.000294 - 0.000907i$. This agrees well with the error predicted by δk.

The eigenvalues of $T_n^{(\mathrm{DtN})}$, $n = 0, 1, 2$, we were able to localize correspond to energies $E = k^2$ with smallest imaginary part and hence the longest lifetime.

Thus, in practice, they are the most likely to be observed and hence the most important. We are able to prove that said eigenvalues from the matrix-valued function $T_n^{(\mathrm{Dir})}$ corresponding to the particle in a box model were already excellent approximations to eigenvalues of $T_n^{(\mathrm{DtN})}$, so the behavior of a system whose potential is well-modelled with the parameters used in constructing $T_n^{(\mathrm{DtN})}$ could be acceptably analyzed using a particle in a box model.

5.5.2 Rational approximation to the DtN map

The information given by the particle in a box model is rather unsatisfying. A handful of good approximations were produced for each n, but from the pseudospectral plot in Figure 5.3 we see that there are plenty of eigenvalues not being captured by eigenvalues of $T_n^{(\mathrm{Dir})}$, including those with small imaginary parts and thus long lifetimes. In addition, we may want to know about certain resonances that are not nearly on the real line or are far from the origin. In this case, the particle in a box model cannot give good approximations, since it predicts real resonances only.

Furthermore, although we obtained good agreement between the particle in a box model and the quantum tunneling model from the previous section, the parameters for the quantum tunneling potential were arbitrarily chosen. Based on work by others [HP96], if we continue to use a length scale where $R = 1$, then it turns out that $V_0 \approx 1204$ and $w \approx 0.02$ are more realistic parameter choices for the potential (5.3) (see Appendix D for a derivation). In this section, denote by $T_n^{(\mathrm{DtN})}$ the matrix-valued function (5.12) constructed by using the more realistic parameters. Unfortunately, the particle in a box model does not agree as well

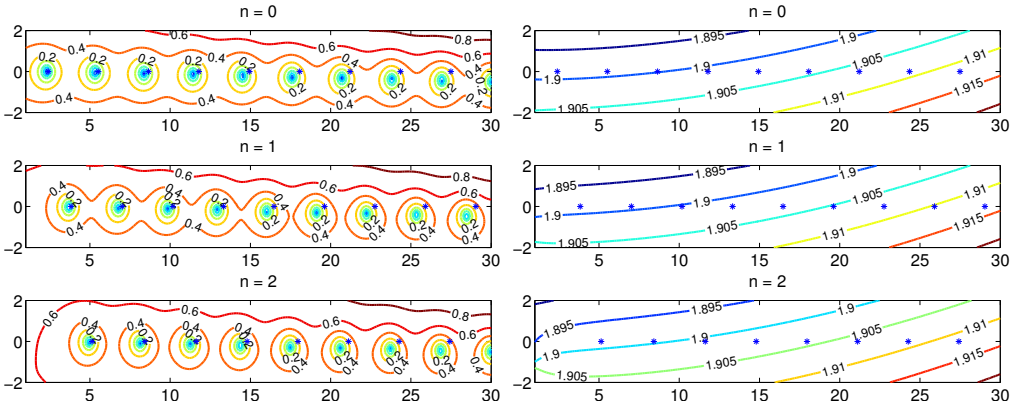

Figure 5.5: Pseudospectra (left) for realistic quantum tunneling model and error (right) between that model and the particle in a box.

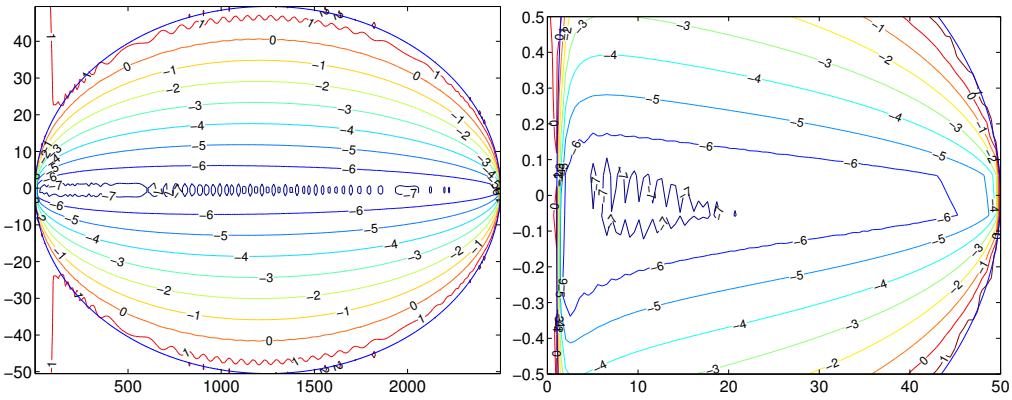

Figure 5.6: The ellipse used to define the rational approximation and contour plot of $\log_{10}|f(z) - f_n(\sqrt{z}, R)|$ for $n = 0$ (left). The square root of the ellipse and contour plot of $\log_{10}\|T_0^{(\mathrm{DtN})}(k) - T_0^{(\mathrm{rat})}(k)\|$ (right), where $T_0^{(\mathrm{rat})}(k)$ is defined according to (5.14).

with this model (see the Figure 5.5 analogous to Figure 5.3), so we are forced to turn elsewhere for eigenvalue approximations.

Let us use a rational approximation to $T_n^{(\mathrm{DtN})}$ to predict its eigenvalues near $[0, 50]$, corresponding to resonance energies $E = k^2$ near $[0, 2500]$. For index $n = 0$ let us define a rational approximation $f(z)$ to $f_n(\sqrt{z}, R)$ according to Proposition 5.1, with Ω the ellipse centered at $1250 - 0.5i$ with semi-major and -minor axis lengths 1249 and 50, and $N = 500$. Following the notation in Proposition 5.1 and (5.14), the rational approximation to $T_n^{(\mathrm{DtN})}$ will be denoted by $T_n^{(\mathrm{rat})}$ and its

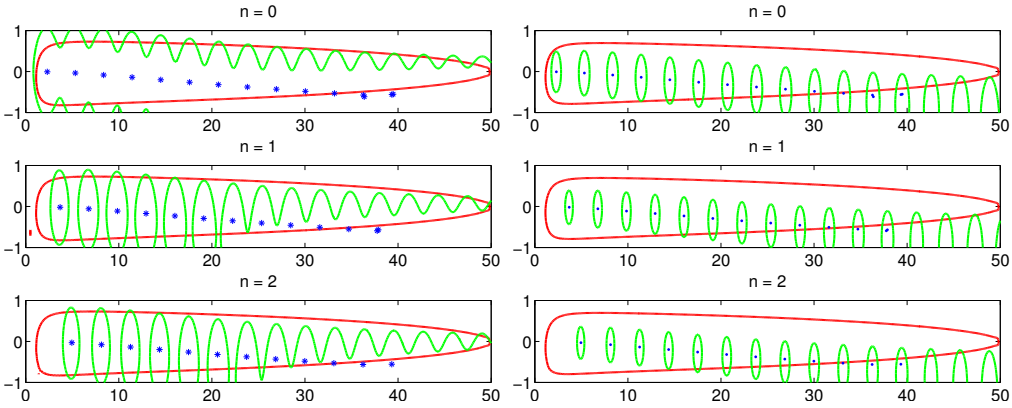

Figure 5.7: Level curves of $\log_{10} \sigma_{\min} T_n^{(\mathrm{DtN})}(k)$ (green) and $\log_{10} \|T_n^{(\mathrm{DtN})}(k) - T_n^{(\mathrm{rat})}(k)\|$ (red) at -1 (left) and -1.3 (right). Compare with Figure 5.6 (right).

linearization will be denoted by $\left(T_{\mathrm{full}}^{(\mathrm{rat})}\right)_n$.

See Figure 5.6 (left) for the $\log_{10}$ error between $f(z)$ and $f_0(\sqrt{z}, R)$ over Ω. Note that $|f(z) - f_0(\sqrt{z}, R)| = \|T_n^{(\mathrm{DtN})}(\sqrt{z}) - T_n^{(\mathrm{rat})}(\sqrt{z})\|_2$ since $T_n^{(\mathrm{DtN})}$ and $T_n^{(\mathrm{rat})}$ differ only in one entry. The accuracy of the rational approximation to $f_0(\sqrt{z}, R)$ suggests that whatever eigenvalues of the pencil $((A_{\mathrm{full}}^{(\mathrm{rat})})_0, (B_{\mathrm{full}}^{(\mathrm{rat})})_0)$ are found near $[0, 2500]$ will be very accurate resonance approximations. Once again, we can use localization results and sensitivity analysis to confirm this. Referring to Figure 5.7 (right), we can see that several of the plotted eigenvalues of $T_n^{(\mathrm{rat})}$ are surrounded by a component (green) of the $10^{-1.2}$-pseudospectrum of $T_n^{(\mathrm{DtN})}$ that fits nicely into the region where $\|T_n^{(\mathrm{DtN})}(k) - T_n^{(\mathrm{rat})}(k)\|_2 < 10^{-1.2}$ (red). By Theorem 2.5, each one that lies entirely within the region where $\|T_n^{(\mathrm{DtN})}(k) - T_n^{(\mathrm{rat})}(k)\|_2 < 10^{-1.2}$ contains exactly one eigenvalue of $T_n^{(\mathrm{DtN})}$ as well. Much tighter localization regions for these eigenvalues of $T_n^{(\mathrm{DtN})}$ can be obtained by looking at ε-pseudospectrum level curves for smaller ε. It should be noted that we can increase the region of accuracy by increasing the number N of poles in the rational approximation f, thus localizing more eigenvalues of $T_n^{(\mathrm{DtN})}$. Alternatively, we can create a region of accuracy that slants down at right, following the direction of the eigenvalues

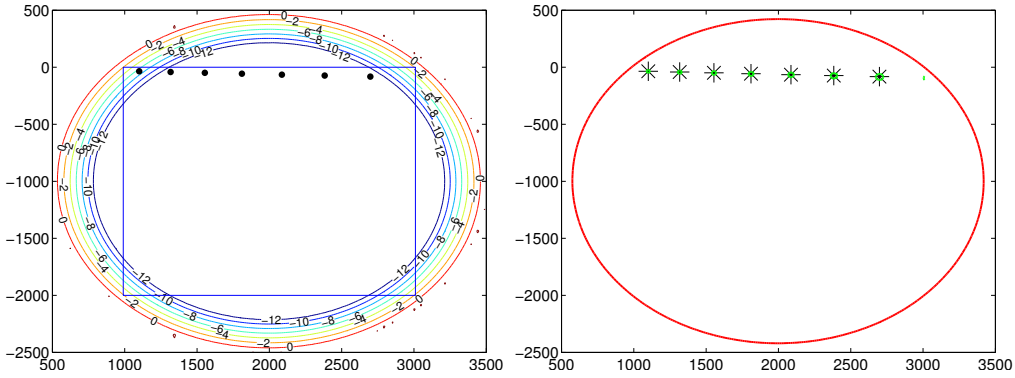

Figure 5.8: Rectangular region of interest, contours of $\log_{10}\|T_n^{(\mathrm{DtN})}(\sqrt{E}) - T_n^{(\mathrm{rat})}(\sqrt{E})\|_2$, and eigenvalues of $T_n^{(\mathrm{rat})}(\sqrt{E})$, all for $n = 0$ (left). The contours $\log_{10}\|T_n^{(\mathrm{DtN})}(\sqrt{E}) - T_n^{(\mathrm{rat})}(\sqrt{E})\|_2 = -1.8$ (red) and $\log_{10}\left(\sigma_{\min}\left(T_n^{(\mathrm{DtN})}(\sqrt{E})\right)\right) = -1.8$ (green), and the eigenvalues of $T_n^{(\mathrm{rat})}(\sqrt{E})$ in the rectangle of interest (right).

of $T_n^{(\mathrm{rat})}$ indicated in Figure 5.7. An angled ellipse is used that way in [BH13, Fig. 6.10].

There is no reason we are restricted to looking near the real line–we can apply the same technique to find resonances far from the real line as well. Suppose we are interested in knowing about resonance energies in another region, say, the rectangle $[990, 3010] \times [-2000, 0] \subset \mathbb{C}$ (note that now we are looking for the energies $E = k^2$, not the parameters k). We used 150 mesh points on a circle of radius 1500 centered at $2000 - 1000i$ to create the rational approximation $T_0^{(\mathrm{rat})}$ whose $\log_{10}$ 2-norm error with $T_0^{(\mathrm{DtN})}$ is pictured in Figure 5.8 (left). Also in Figure 5.8 (left), we plot resonance energy approximations computed from the eigenvalues of $\left(T_{\mathrm{full}}^{(\mathrm{rat})}\right)_0$, which is the linearization of the rational approximation $T_0^{(\mathrm{rat})}$ to $T_0^{(\mathrm{DtN})}$. On the right, we can see that the resonance approximations computed from $\left(T_{\mathrm{full}}^{(\mathrm{rat})}\right)_0$ are each in their own connected component of the $10^{-1.8}$-pseudospectrum of $T_n^{(\mathrm{DtN})}$, and all lie within the region where $\|T_n^{(\mathrm{DtN})}(\sqrt{E}) - T_n^{(\mathrm{rat})}(\sqrt{E})\|_2 < 10^{-1.8}$. Therefore, each of those connected components contains exactly one eigenvalue of $T_n^{(\mathrm{DtN})}$ by Theorem 2.5. With the reso-

nances of $T_0^{(\mathrm{DtN})}$ in the chosen rectangle localized, we are now free to compute them using standard techniques described in Step 5 of Section 2.3.1.

5.6 Conclusion

We have presented a framework for computing definitive bounds on the difference between quantum corral resonances predicted by different discretized elastic scattering models with axisymmetric potentials. Along with it is a method of resonance computation guided by approximation and localization as described in Section 2.3.1. We have also discussed sources of error due to the discretization and offered first-order perturbation analysis. The framework is flexible enough to be used with finite elements rather than spectral collocation, and in theory is easily extended to the general two-dimensional case, although the computational challenges increase significantly. Once the computational obstacles are overcome, however, this will be the only two-dimensional resonance computation framework of which the author is aware that permits concrete error analysis.

CHAPTER 6

CONCLUSION

We have presented tools to make nonlinear eigenvalue problems easier. Various strategies for accomplishing this were illustrated on a collection of examples. In addition, we have used these tools in other ways, obtaining concrete bounds on transient growth for linear delay differential equations via generalized pseudospectra; appealing to our generalization of Gershgorin's theorem in an analytical derivation of a bracketing interval for the positive, real eigenvalue of the `fiber` problem from [BHM+13]; and invoking the same theorem in the proof in Section 5.4, where we showed that the eigenvalues of a sequence of matrix-valued functions eventually leave any bounded region of interest. The author hopes that others will find further creative ways to employ these flexible and practical theorems, and that the exposition given here will serve as a helpful resource.

APPENDIX A

DELTA POTENTIALS

The time-independent, one-dimensional Schrödinger equation is

$$\left(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right)\psi(x) = E\psi(x) \tag{A.1}$$

where $\hbar$ is the reduced Planck constant, $V(x)$ is called the potential function, E is energy, and m is the reduced mass. As a facile example, m could be the mass of an electron with energy E (insofar as an electron could be confined to a line), in which case $\int_a^b |\psi(x)|^2\,dx$ would represent the probability of finding the electron in the interval $[a, b]$. A famous first example for quantum students is the one-dimensional "particle in a box" potential $V(x) = 0$ in $[0, a]$ and $V(x) = \infty$ otherwise, in which case the particle in the box (say, the electron) is allowed only certain energies E_n and corresponding wavefunctions (bound states) $\psi_n(x)$ [Gri05]:

$$E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{n^2\pi^2\hbar^2}{2ma^2}, \qquad \psi_n(x) = \sqrt{\frac{2}{a}}\sin\left(\frac{n\pi}{a}x\right) \qquad (n = 1, 2, ...). \tag{A.2}$$

If ψ is a solution to (A.1), then the time-dependent Schrödinger equation dictates that $e^{-iEt/\hbar}\psi(x)$ is its evolution as a function of time. In a region where $V(x) = 0$, ψ is a linear combination of e^{ikx} and e^{-ikx}, where $E = \hbar^2k^2/(2m)$ as in (A.2). Now, $e^{-iEt/\hbar}e^{ikx} = e^{i(kx - Et/\hbar)}$. If $k > 0$, then in order for $kx - Et/\hbar$ to stay constant as time increases, x must increase, and hence e^{ikx} corresponds to a wave traveling to the right. Similarly, e^{-ikx} travels to the left.

In the rest of this appendix we will use the following scaled version of (A.1) without the presence of those physical constants:

$$\left(-\frac{\partial^2}{\partial x^2} + V(x)\right)\psi(x) = k^2\psi \tag{A.3}$$

(where the k corresponds to k_n above). We will call k^2 the energy from now on. Furthermore, we will no longer be interested in finding bound states, but instead will consider scattering problems and scattering resonances. In a scattering problem, k may range over a continuous set and the solution ψ is not required to be square integrable.

It is helpful to break $\psi(x)$ into a scattered part (defined to be left traveling at $-\infty$ and right traveling at ∞) and an incident part–this terminology comes from the idea of scattering experiments where a stream of particles is fired at an obstacle. For a compactly-supported potential $V(x)$, it is clear on physical grounds that the incident wave ψ_{inc} satisfies $-\psi''_{\text{inc}}(x) = k^2\psi_{\text{inc}}$. And from the previous paragraph, the scattered wave must be a multiple of e^{-ikx} to the left of the potential support and must be a multiple of e^{ikx} to the right. Writing $\psi = \psi_{\text{scatt}} + \psi_{\text{inc}}$, this characterizes the scattered wave with the differential equation $\left(-\partial^2/\partial x^2 + V(x) - k^2\right)\psi_{\text{scatt}} = -V(x)\psi_{\text{inc}}$ and the above boundary conditions. Resonances correspond to values of k^2 such that a solution exists even when we set $\psi_{\text{inc}} = 0$.

Another first example for physics students is a potential function made up of weighted delta functions (see [Gri05] for a discussion), say $V(x) = \sum_{n=1}^{N} \alpha_n \delta(x - x_n)$ where $x_1 < x_2 < ... < x_N$. For a given energy k^2, the solution ψ must look like some linear combination of e^{ikx} and e^{-ikx} between each of the delta functions (different linear combinations in each subinterval). The solution ψ must be con-

tinuous everywhere, and using the definition of delta function

$$\lim_{\varepsilon\to 0^+}\int_{x_n-\varepsilon}^{x_n+\varepsilon}\left(-\frac{\partial^2}{\partial x^2}+V(x)\right)\psi(x)\,dx=\lim_{\varepsilon\to 0^+}\int_{x_n-\varepsilon}^{x_n+\varepsilon}k^2\psi(x)\,dx \tag{A.4}$$

$$\Leftrightarrow \lim_{\varepsilon\to 0^+}\int_{x_n-\varepsilon}^{x_n+\varepsilon}\psi''(x)\,dx=\lim_{\varepsilon\to 0^+}\int_{x_n-\varepsilon}^{x_n+\varepsilon}V(x)\psi(x)\,dx \qquad \text{(continuity of the wavefunction)} \tag{A.5}$$

$$\Leftrightarrow \lim_{\varepsilon\to 0^+}\psi'(x_n+\varepsilon)-\psi'(x_n-\varepsilon)=\alpha_n\psi(x_n). \tag{A.6}$$

For concreteness, let us take the case of two delta functions, one placed at zero and the other at $x = p$

$$V(x)=\alpha_1\delta(x)+\alpha_2\delta(x-p) \tag{A.7}$$

and write the solution as

$$\psi(x)=\begin{cases}\psi_1(x)=A_1e^{ikx}+B_1e^{-ikx}, & x<0\\ \psi_2(x)=A_2e^{ikx}+B_2e^{-ikx}, & 0<x<p\\ \psi_3(x)=A_3e^{ikx}+B_3e^{-ikx}, & p<x\end{cases}. \tag{A.8}$$

Then the continuity conditions are $\psi_1(0) = \psi_2(0)$ and $\psi_2(p) = \psi_3(p)$. Writing the derivative jump conditions as $\psi_2'(0)-\psi_1'(0) = \alpha_1(\psi_1(0)+\psi_2(0))/2$ and $\psi_3'(p)-\psi_2'(p) = \alpha_2(\psi_2(p)+\psi_3(p))/2$, we obtain four conditions

$$\begin{aligned}A_1+B_1-A_2-B_2&=0\\ A_1(\alpha_1/2+ik)+B_1(\alpha_1/2-ik)+A_2(\alpha_1/2-ik)+B_2(\alpha_1/2+ik)&=0\\ A_2e^{ikp}+B_2e^{-ikp}-A_3e^{ikp}-B_3e^{-ikp}&=0\\ A_2(\alpha_2/2+ik)e^{ikp}+B_2(\alpha_2/2-ik)e^{-ikp}+A_3(\alpha_2/2-ik)e^{ikp}+B_3(\alpha_2/2+ik)e^{-ikp}&=0\end{aligned} \tag{A.9}$$

which is the matrix equation

$$\begin{bmatrix} 1 & 1 & -1 & -1 & 0 & 0 \\ \frac{\alpha_1}{2}+ik & \frac{\alpha_1}{2}-ik & \frac{\alpha_1}{2}-ik & \frac{\alpha_1}{2}+ik & 0 & 0 \\ 0 & 0 & e^{ikp} & e^{-ikp} & -e^{ikp} & -e^{-ikp} \\ 0 & 0 & \left(\frac{\alpha_2}{2}+ik\right)e^{ikp} & \left(\frac{\alpha_2}{2}-ik\right)e^{-ikp} & \left(\frac{\alpha_2}{2}-ik\right)e^{ikp} & \left(\frac{\alpha_2}{2}+ik\right)e^{-ikp} \end{bmatrix} \begin{bmatrix} A_1 \\ B_1 \\ A_2 \\ B_2 \\ A_3 \\ B_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{A.10}$$

According to [Gri05], a typical physical experiment corresponds to $B_3 = 0$ and A_1 set by the experimenter, leading to

$$\begin{bmatrix} 1 & -1 & -1 & 0 \\ \frac{\alpha_1}{2}-ik & \frac{\alpha_1}{2}-ik & \frac{\alpha_1}{2}+ik & 0 \\ 0 & e^{ikp} & e^{-ikp} & -e^{ikp} \\ 0 & \left(\frac{\alpha_2}{2}-ik\right)e^{-ikp} & \left(\frac{\alpha_2}{2}-ik\right)e^{ikp} & \left(\frac{\alpha_2}{2}+ik\right)e^{-ikp} \end{bmatrix} \begin{bmatrix} B_1 \\ A_2 \\ B_2 \\ A_3 \end{bmatrix} = \begin{bmatrix} -A_1 \\ -A_1\left(\frac{\alpha_1}{2}+ik\right) \\ 0 \\ 0 \end{bmatrix}. \tag{A.11}$$

Equivalently, we can consider the scattered wave picture with incident wave $\psi_{\text{inc}}(x) = A_1 e^{ikx}$. If we reuse notation, defining

$$\psi_{\text{scatt}}(x) = \begin{cases} B_1 e^{-ikx}, & x < 0 \\ A_2 e^{ikx} + B_2 e^{-ikx}, & 0 < x < p \\ A_3 e^{ikx}, & p < x \end{cases} \tag{A.12}$$

and $\psi(x) = \psi_{\text{scatt}}(x) + A_1 e^{ikx}$, then the total wave ψ is

$$\psi(x) = \begin{cases} A_1 e^{ikx} + B_1 e^{-ikx}, & x < 0 \\ \tilde{A}_2 e^{ikx} + B_2 e^{-ikx}, & 0 < x < p \\ \tilde{A}_3 e^{ikx}, & p < x \end{cases}, \tag{A.13}$$

where $\tilde{A}_2 = A_2 + A_1$ and $\tilde{A}_3 = A_3 + A_1$. Applying the continuity conditions as before, we end up with

$$\begin{bmatrix} 1 & -1 & -1 & 0 \\ \frac{\alpha_1}{2} - ik & \frac{\alpha_1}{2} - ik & \frac{\alpha_1}{2} + ik & 0 \\ 0 & e^{ikp} & e^{-ikp} & -e^{ikp} \\ 0 & \left(\frac{\alpha_2}{2} - ik\right) e^{-ikp} & \left(\frac{\alpha_2}{2} - ik\right) e^{ikp} & \left(\frac{\alpha_2}{2} + ik\right) e^{-ikp} \end{bmatrix} \begin{bmatrix} B_1 \\ \tilde{A}_2 \\ B_2 \\ \tilde{A}_3 \end{bmatrix} = \begin{bmatrix} -A_1 \\ -A_1\left(\frac{\alpha_1}{2} + ik\right) \\ 0 \\ 0 \end{bmatrix}. \quad \text{(A.14)}$$

Putting $\psi_{\text{inc}} = 0$ is to put $A_1 = 0$, yielding

$$\begin{bmatrix} 1 & -1 & -1 & 0 \\ \frac{\alpha_1}{2} - ik & \frac{\alpha_1}{2} - ik & \frac{\alpha_1}{2} + ik & 0 \\ 0 & e^{ikp} & e^{-ikp} & -e^{ikp} \\ 0 & \left(\frac{\alpha_2}{2} - ik\right) e^{-ikp} & \left(\frac{\alpha_2}{2} - ik\right) e^{ikp} & \left(\frac{\alpha_2}{2} + ik\right) e^{-ikp} \end{bmatrix} \begin{bmatrix} B_1 \\ A_2 \\ B_2 \\ A_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad \text{(A.15)}$$

This characterizes the resonances as the energies k^2 such that this matrix equation admits a nonzero solution, i.e., such that the matrix is singular.

APPENDIX B

CHANGE OF VARIABLE MAKING GUN POLYNOMIAL

The gun problem from [BHM+13] is of the form

$$F(\lambda) = K - \lambda M + i\sqrt{\lambda} W_1 + i\sqrt{\lambda - \sigma_2^2} W_2, \tag{B.1}$$

where according to [Lia07, p. 59] the eigenvalues of interest are a subset of those with $\sqrt{\lambda}$ near and to the right of 146.71, where the principal branch of the square root is used. In this appendix, we show how the problem of finding the eigenvalues of F can be reformulated as a polynomial eigenvalue problem.

First define $z = \sqrt{\lambda}$, taking the principal branch. This means that the variable z represents a value in the union of the open right half-plane and the ray $i(0, \infty)$. Using this, we have

$$F(\lambda) = K - z^2 M + iz W_1 + iz\sqrt{1 - (\sigma_2/z)^2} W_2, \qquad \lambda = z^2. \tag{B.2}$$

This suggests we put $\sigma_2/z = \cos(\theta)$ so that $\sqrt{1 - (\sigma_2/z)^2}$ equals $\sin(\theta)$. Defining $w = e^{i\theta}$, $0 \le \theta < 2\pi$, such a change of variable is equivalent to

$$\frac{2\sigma_2}{z} = w + w^{-1}. \tag{B.3}$$

Rewriting as a quadratic in w, it follows that for a fixed value of z there are two values of w that satisfy this equation. Therefore we have not lost any eigenvalues of (B.2) by rewriting in terms of w. In terms of the θ variable, (B.1) becomes

$$\cos^2\theta\, F(\lambda) = K\cos^2\theta - \sigma_2^2 M + i\sigma_2 \cos\theta W_1 + i\sigma_2 \cos\theta \sin\theta W_2, \qquad \lambda = (\sigma_2/\cos\theta)^2. \tag{B.4}$$

Next, observing that $2i\sin\theta = w - w^{-1}$, in terms of the w variable

$$\cos^2\theta\, F(\lambda) \tag{B.5}$$

$$= K\left(\frac{w+w^{-1}}{2}\right)^2 - \sigma_2^2 M + i\sigma_2\left(\frac{w+w^{-1}}{2}\right)W_1 + i\sigma_2\left(\frac{w+w^{-1}}{2}\right)\left(\frac{w-w^{-1}}{2i}\right)W_2 \tag{B.6}$$

$$= \frac{1}{4}K\left(w^2+2+w^{-2}\right) - \sigma_2^2 M + \frac{1}{2}i\sigma_2\left(w+w^{-1}\right)W_1 + \frac{1}{4}\sigma_2\left(w^2-w^{-2}\right)W_2, \tag{B.7}$$

$$\lambda = \left(\frac{2\sigma_2}{w+w^{-1}}\right)^2. \tag{B.8}$$

Multiplying by $4w^2$ gives

$$4w^2\cos^2\theta\, F(\lambda) \tag{B.9}$$

$$= K\left(w^4+2w^2+1\right) - 4w^2\sigma_2^2 M + 2i\sigma_2\left(w^3+w\right)W_1 + \sigma_2\left(w^4-1\right)W_2 \tag{B.10}$$

$$= (K+\sigma_2 W_2)\,w^4 + (2i\sigma_2 W_1)\,w^3 + \left(2K - 4\sigma_2^2 M\right)w^2 + (2i\sigma_2 W_1)\,w + (K-\sigma_2 W_2) \tag{B.11}$$

$$:= P(w). \tag{B.12}$$

The matrix-valued function P is a matrix polynomial in w. Noticing that $4w^2\cos^2\theta = (w^2+1)^2$, it follows that $P(w_0)$ is singular at a particular value w_0 if and only if $\lambda_0 = \left(2\sigma_2/(w_0 + w_0^{-1})\right)^2$ is an eigenvalue of F or $w_0 = \pm i$. It is worth noting that $w = \pm i$ corresponds to $\theta = \pm\pi/2$, or $\cos\theta = 0$, which cannot occur for finite values of z (or λ). So the spectrum of F is completely characterized by the spectrum of P.

Since the matrices in the definition of F and P are so large, the natural thing to do is use a sparse solver to find eigenvalues of P near some point of interest. Now, if we are interested in z near but to the right of $z_0 = 146.71$, there are two values of w that satisfy the change of variable equation $z_0 = 2\sigma_2/(w+w^{-1})$, namely

$$w_\pm = \frac{\sigma_2}{z_0} \pm \sqrt{\left(\frac{\sigma_2}{z_0}\right)^2 - 1}. \tag{B.13}$$

Therefore to find eigenvalues of F with $\lambda^{1/2}$ near z_0, we should look for eigenvalues of P near both w_+ and w_-. For the given problem parameters, $w_\pm$ are $\approx 0.7421 \pm 0.6703i$, which are sufficiently close together (and sufficiently separated from the rest of the spectrum) that it suffices to look near only one of them. This stays true if z_0 is taken further to the right, since such a choice decreases σ_2/z_0 towards zero and thus makes the difference between $w_\pm$ approach 2.

APPENDIX C

THE DIRICHLET-TO-NEUMANN MAP

In Section 5.4.1, first-order perturbation theory for 2D scattering resonances made use of a Dirichlet-to-Neumann (DtN) map $f(k,R)$ and some of its properties. Most importantly, we used

$$\int_{\partial\Omega} uf(k,R)v\,dS = \int_{\partial\Omega} vf(k,R)u\,dS \tag{C.1}$$

for $\Omega = B(0,R)$ and any $u, v \in H^1(\Omega)$, and the proof is provided in this section. For simplicity of notation, we use $B(k)$ instead of $f(k,R)$ in all that follows.

Before starting, note that the appearance of $B(k)v$ in the boundary integral in (C.1) represents the DtN map $B(k)$ applied to the trace Tv of v, i.e., $v|_{\partial\Omega}$ in case $v \in C^\infty(\bar{\Omega})$. We will postpone discussion of traces and other technical matters, and for now give a proof of (C.1) under a set of restrictive assumptions.

Lemma C.1. *Suppose $u, v \in C^\infty(\bar{\Omega})$ with $f := u|_{\partial\Omega}$ and $g := |_{\partial\Omega}$. If $B(k)f, B(k)g \in L^2(\partial\Omega)$, then (C.1) holds.*

Proof. First of all, recall that Parseval's theorem says if $a, b \in L^2(\partial\Omega)$, then $\int_0^{2\pi} a(\theta)\overline{b(\theta)}\,d\theta = 2\pi\sum_n a_n\overline{b_n}$, where $a(\theta) = \sum_n a_n e^{in\theta}$ and $b(\theta) = \sum_n b_n e^{in\theta}$. Therefore

$$2\pi R\sum_n a_n\overline{b_n} = \int_0^{2\pi} a(\theta)\overline{b(\theta)}\,Rd\theta = \int_{\partial\Omega} a\bar{b}\,dS. \tag{C.2}$$

Second, if $B(k)$ is the DtN map acting on functions on $\partial\Omega$, then $f \mapsto B(k)f$ is formally defined by

$$B(k)f = \sum_n B(k)_n f_n e^{in\theta}, \qquad B(k)_n = k\frac{\left(H_n^{(1)}\right)'(kR)}{H_n^{(1)}(kR)} \tag{C.3}$$

where $H_n^{(1)}(x)$ is the first-kind (i.e. outgoing) Hankel function equal to $J_n(x) + iY_n(x)$. From standard properties of integer-order Bessel functions, we have $H_{-n}^{(1)} = (-1)^n H_n^{(1)}$, $\left(H_n^{(1)}\right)' = \left[H_{n-1}^{(1)} - H_{n+1}^{(1)}\right]/2$, and $\left(H_{-n}^{(1)}\right)' = (-1)^n \left(H_n^{(1)}\right)'$ from which it follows that

$$B(k)_{-n} = B(k)_n. \tag{C.4}$$

Now define $G = \overline{B(k)g} \in L^2(\partial\Omega)$ and $F = \overline{B(k)f} \in L^2(\partial\Omega)$ so that $B(k)g = \bar{G}$ and $B(k)f = \bar{F}$. Then for $G(\theta) = \sum_n G_n e^{in\theta}$,

$$G_n = \frac{1}{2\pi}\int_0^{2\pi} G(\theta)e^{-in\theta}\,d\theta = \frac{1}{2\pi}\int_0^{2\pi} \overline{B(k)ge^{in\theta}}\,d\theta = \overline{\frac{1}{2\pi}\int_0^{2\pi} B(k)ge^{in\theta}\,d\theta} \tag{C.5}$$

$$= \overline{(B(k)g)_{-n}}. \tag{C.6}$$

Therefore $\overline{G_n} = (B(k)g)_{-n}$, which is just $B(k)_{-n}g_{-n}$, and similarly for F_n. Since f and g are obviously in $L^2(\partial\Omega)$ because they are smooth, and $B(k)f, B(k)g \in L^2(\partial\Omega)$ by assumption, we have

$$\int_{\partial\Omega} fB(k)g\,dS = \int_{\partial\Omega} f\bar{G}\,dS \tag{C.7}$$

$$= 2\pi R\sum_n f_n\overline{G_n} \qquad \text{(from (C.2))} \tag{C.8}$$

$$= 2\pi R\sum_n f_n\,(B(k)g)_{-n} \tag{C.9}$$

$$= 2\pi R\sum_n f_n B(k)_{-n}g_{-n} \tag{C.10}$$

$$= 2\pi R\sum_n f_{-n}B(k)_n g_n \qquad \text{(reverse order of summation)} \tag{C.11}$$

$$= 2\pi R\sum_n g_n B(k)_{-n}f_{-n} \qquad \text{(from (C.4))} \tag{C.12}$$

$$= 2\pi R\sum_n g_n\,(B(k)f)_{-n} \tag{C.13}$$

$$= 2\pi R \sum_n g_n \overline{F_n} \tag{C.14}$$

$$= \int_{\partial\Omega} g\bar{F}\, dS \qquad \text{(from (C.2))} \tag{C.15}$$

$$= \int_{\partial\Omega} gB(k)f\, dS. \tag{C.16}$$

Therefore (C.1) holds under the assumptions of this lemma. This completes the proof. □

It's simple to show that $u \in C^\infty(\bar{\Omega})$ implies $B(k)f \in L^2(\partial\Omega)$ once we have some estimates for $B(k)_n$. The next lemma includes these.

Lemma C.2.

$$B(k)_n \to -\frac{|n|}{R} + \frac{1}{|n|-1}\,\frac{k^2R}{4} \qquad as \qquad |n| \to \infty \tag{C.17}$$

and

$$\frac{d}{dk}B(k)_n \to \frac{1}{|n|-1}\,\frac{3kR}{4} + \frac{1}{(n-1)^2(n-2)}\,\frac{(kR)^3}{2^4} \qquad as \qquad |n| \to \infty. \tag{C.18}$$

Proof. By (C.4), it suffices to consider $n > 0$. According to [AS70, §9.3], for fixed x and $n \to \infty$ we have

$$J_n(x) \approx \frac{1}{\Gamma(n+1)}\left(\frac{x}{2}\right)^n \qquad \text{and} \qquad Y_n(x) \approx -\frac{\Gamma(n)}{\pi}\left(\frac{2}{x}\right)^n. \tag{C.19}$$

Due to the speedy decrease of $J_n(x)$ as $n \to \infty$, it is clear that $H_n \approx iY_n$ and $H_n' \approx iY_n'$ for n large enough. Hence,

$$\frac{H_n'(x)}{H_n(x)} \approx \frac{Y_{n-1}(x) - Y_{n+1}(x)}{2Y_n(x)} \approx \frac{\Gamma(n-1)\left(\frac{2}{x}\right)^{n-1} - \Gamma(n+1)\left(\frac{2}{x}\right)^{n+1}}{2\Gamma(n)\left(\frac{2}{x}\right)^n} \tag{C.20}$$

$$= \frac{1}{2}\left(\frac{1}{n-1}\left(\frac{2}{x}\right)^{-1} - n\left(\frac{2}{x}\right)\right) = -\frac{n}{x} + \frac{1}{n-1}\,\frac{x}{4}. \tag{C.21}$$

Thus,

$$B(k)_n = k\frac{H_n'(kR)}{H_n(kR)} \approx -\frac{n}{R} + \frac{1}{n-1}\,\frac{k^2R}{4} \tag{C.22}$$

for large enough n.

Now we consider $\frac{d}{dk}B(k)_n$. This is

$$\frac{d}{dk}B(k)_n = \frac{1}{k}B(k)_n + kR\frac{H_n''(kR)H_n(kR) - H_n'(kR)^2}{H_n(kR)^2}. \tag{C.23}$$

It is again clear that we may take $n > 0$ with no loss of generality. Using asymptotic expressions for H_n as above, we have

$$H_n(x) \approx -\frac{i}{\pi}(n-1)!\left(\frac{2}{x}\right)^n \tag{C.24}$$

$$H_n(x)^2 \approx -\frac{1}{\pi^2}(n-1)!^2\left(\frac{2}{x}\right)^{2n} \tag{C.25}$$

$$H_n'(x) = \frac{1}{2}(H_{n-1}(x) - H_{n+1}(x)) \tag{C.26}$$

$$\approx -\frac{i}{2\pi}\left(\frac{2}{x}\right)^n\left((n-2)!\left(\frac{2}{x}\right)^{-1} - n!\left(\frac{2}{x}\right)\right) \tag{C.27}$$

$$H_n''(x) = \frac{1}{4}(H_{n-2}(x) - 2H_n(x) + H_{n+2}(x)) \tag{C.28}$$

$$\approx -\frac{i}{4\pi}\left(\frac{2}{x}\right)^n\left((n-3)!\left(\frac{2}{x}\right)^{-2} - 2(n-1)! + (n+1)!\left(\frac{2}{x}\right)^2\right). \tag{C.29}$$

After simplifying the factorials, we obtain $H_n''(x)H_n(x) - H_n'(x)^2 \approx$

$$-\frac{1}{4\pi^2}\left(\frac{2}{x}\right)^{2n}\left[(n-3)!^2(n-2)\left(\frac{2}{x}\right)^{-2} + 2(n-2)!^2(n-1) + (n-1)!^2 n\left(\frac{2}{x}\right)^2\right] \tag{C.30}$$

showing that

$$\frac{H_n''(x)H_n(x) - H_n'(x)^2}{H_n(x)^2} \approx \frac{1}{4}\left[\frac{1}{(n-1)^2(n-2)}\left(\frac{x}{2}\right)^2 + 2\frac{1}{n-1} + n\left(\frac{2}{kR}\right)^2\right]. \tag{C.31}$$

Hence, for large n,

$$\frac{d}{dk}B(k)_n \approx -\frac{n}{kR} + \frac{1}{n-1}\,\frac{kR}{4} + \frac{kR}{4}\left[\frac{1}{(n-1)^2(n-2)}\left(\frac{kR}{2}\right)^2 + 2\frac{1}{n-1} + n\left(\frac{2}{kR}\right)^2\right] \tag{C.32}$$

$$= \frac{1}{n-1}\,\frac{3kR}{4} + \frac{1}{(n-1)^2(n-2)}\,\frac{(kR)^3}{2^4}. \tag{C.33}$$

□

With these estimates in hand, we can simplify the assumptions in Lemma C.1.

Lemma C.3. *If $u, v \in C^\infty(\bar{\Omega})$, then (C.1) holds.*

Proof. From $u \in C^\infty(\bar{\Omega}$, we see that $f \in C^\infty(\partial\Omega)$, and so f and f' are both in $L^2(\partial\Omega)$. Furthermore, by integration by parts

$$(f')_n = \frac{1}{2\pi}\int_0^{2\pi} f'(\theta)e^{-in\theta}\,d\theta \tag{C.34}$$

$$= \frac{1}{2\pi}\left[f(\theta)e^{-in\theta}\Big|_0^{2\pi} - \int_0^{2\pi} f(\theta)(-in)e^{-in\theta}\,d\theta\right] \tag{C.35}$$

$$= inf_n. \tag{C.36}$$

By (C.2), $(nf_n) \in \ell^2$. It follows from the asymptotics for $B(k)_n$ in Lemma C.2 that $(B(k)_n f_n) \in \ell^2$. By (C.2) once again, $\|B(k)f\|_{L^2(\partial\Omega)} < \infty$. Similarly, $B(k)g \in L^2(\partial\Omega)$. Therefore the assumptions of Lemma C.1 are satisfied, and hence (C.1) holds for $u, v \in C^\infty(\bar{\Omega})$. □

Now, $C^\infty(\bar{\Omega})$ is dense in $H^1(\Omega)$ [Eva98, §5], so if the mapping $u \times v \mapsto \int_{\partial\Omega} uB(k)v\,dS$ is continuous on $H^1(\Omega)\times H^1(\Omega)$, then (C.1) holds for all $u, v \in H^1(\Omega)$ by Lemma C.3. We do need a few technical results now, starting with traces and the trace theorem.

Let T be the trace operator whose restriction to $C^\infty(\bar{\Omega})$ acts as $Tu = u|_{\partial\Omega}$. The first lemma is a version of the trace theorem (see [Wlo87] for a rigorous proof of the general result). In an effort to keep this appendix self-contained, we include the proof for the special case $\Omega = B(0, R)$.

Lemma C.4. *If $u \in H^1(\Omega)$, then*

$$\|(n^{1/2}(Tu)_n)\|_{\ell^2}^2 \le \frac{1}{2\pi}\|\nabla u\|_{L^2(\Omega)}^2. \tag{C.37}$$

Proof. Take $u \in C^\infty(\bar{\Omega})$ real-valued and define $f := Tu$. Then, $f \in C^\infty(\partial\Omega)$, and hence it has Poisson extension

$$v(r,\theta) = \begin{cases} \frac{1}{2\pi}\int_0^{2\pi} P_{r/R}(\theta - t)\, f(t)\, dt, & 0 \le r < R \\ f(\theta), & r = R \end{cases} \tag{C.38}$$

to all of $\bar{\Omega}$, where $P_{r/R}$ is the Poisson kernel:

$$P_{r/R}(\varphi) = \sum_{n=-\infty}^{\infty} \left(\frac{r}{R}\right)^{|n|} e^{in\varphi}. \tag{C.39}$$

The Poisson extension v is known to be harmonic and $C^\infty(\bar{\Omega})$.

Next, we show that $\|\nabla v\|^2_{L^2(\Omega)} = 2\pi\|(n^{1/2} f_n)\|^2_{\ell^2}$. First, observe that since $v \in C^\infty(\bar{\Omega})$, all of f, f' and $T(dv/dr)$ are in $L^2(\partial\Omega)$. Now, the integrand defining v is absolutely convergent, and therefore uniformly convergent, on compact subsets of Ω. Hence, we can integrate term by term to get

$$v = \sum_{n=-\infty}^{\infty} \left(\frac{r}{R}\right)^{|n|} e^{in\theta} f_n. \tag{C.40}$$

For the same reason, we can differentiate this series term by term to get

$$\frac{dv}{dr} = \frac{1}{R}\sum_{n=-\infty}^{\infty} |n| \left(\frac{r}{R}\right)^{|n|-1} e^{in\theta} f_n. \tag{C.41}$$

Hence, the n-th Fourier coefficient of $T(dv/dr)$ is

$$\left(T\left(\frac{dv}{dr}\right)\right)_n = \frac{|n|}{R} f_n. \tag{C.42}$$

Therefore, since $f = \bar{f}$ by the assumption that u is real-valued,

$$\begin{aligned} \int_{\partial\Omega} v\,(\nabla v \cdot \mathbf{n})\, dS &= \int_0^{2\pi} f \left(T\left(\frac{dv}{dr}\right)\right) R d\theta && \text{(C.43)} \\ &= 2\pi R \frac{1}{2\pi}\int_0^{2\pi} \bar{f}\left(T\left(\frac{dv}{dr}\right)\right) d\theta && \text{(C.44)} \\ &= 2\pi R \sum_{n=-\infty}^{\infty} \overline{f_n}\frac{|n|}{R} f_n \qquad \text{(by (C.2))} && \text{(C.45)} \\ &= 2\pi \sum_{n=-\infty}^{\infty} |n|\,|f_n|^2. && \text{(C.46)} \end{aligned}$$

An integration by parts using Green's first identity shows

$$\int_{\partial\Omega} v(\nabla v \cdot \mathbf{n})\, dS = \int_{\Omega} (v\Delta v + |\nabla v|^2)\, dA = \|\nabla v\|^2_{L^2(\Omega)} \tag{C.47}$$

because v is harmonic, and therefore $\|\nabla v\|^2_{L^2(\Omega)} = 2\pi\|(n^{1/2} f_n)\|^2_{\ell^2}$ as desired.

The final major step is to show that $\|\nabla v\|_{L^2(\Omega)} \leq \|\nabla u\|_{L^2(\Omega)}$ (that is, of all smooth functions with the same trace, the harmonic one has smallest H^1 semi-norm). We show this by writing $u = v + (u - v)$ and using integration by parts. Thus we obtain

$$\int_{\Omega} |\nabla u|^2\, dA = \int_{\Omega} |\nabla v|^2\, dA + \int_{\Omega} |\nabla(u - v)|^2\, dA + 2\int_{\Omega} \nabla v \cdot \nabla(u - v)\, dA \tag{C.48}$$

since u and v are real. Applying Green's first identity,

$$\int_{\Omega} \nabla v \cdot \nabla(u - v)\, dA = \int_{\partial\Omega} (u - v)(\nabla v \cdot \mathbf{n})\, dS - \int_{\Omega} (u - v)\Delta v\, dA = 0 \tag{C.49}$$

where $Tu = Tv$ and $\Delta v = 0$ were used. Therefore

$$\int_{\Omega} |\nabla u|^2\, dA = \int_{\Omega} |\nabla v|^2\, dA + \int_{\Omega} |\nabla(u - v)|^2\, dA \geq \int_{\Omega} |\nabla v|^2\, dA \tag{C.50}$$

as required.

In summary, so far we have proved that

$$2\pi \sum_{n=-\infty}^{\infty} |n|\, |f_n|^2 = \|\nabla v\|^2_{L^2(\Omega)} \leq \|\nabla u\|^2_{L^2(\Omega)} \tag{C.51}$$

for all real-valued $u \in C^\infty(\bar{\Omega})$. Now, if u is complex-valued, then the inequality (C.51) holds for both its real part $\mathfrak{Re}u$ and its imaginary part $\mathfrak{Im}u$. Since $|\nabla u|^2 = |\nabla \mathfrak{Re}u|^2 + |\nabla \mathfrak{Im}u|^2$ and $|f_n|^2 = |(\mathfrak{Re}f)_n|^2 + |(\mathfrak{Im}f)_n|^2$, the desired inequality holds for all $u \in C^\infty(\bar{\Omega})$. The density of $C^\infty(\bar{\Omega})$ in $H^1(\Omega)$ shows the inequality holds for all $u \in H^1(\Omega)$, completing the proof. □

Corollary C.1. *If $u \in H^1(\Omega)$, then $Tu \in L^2(\partial\Omega)$.*

Proof. This follows from

$$\|Tu\|^2_{L^2(\partial\Omega)} = 2\pi R \sum_{n=-\infty}^{\infty} |f_n|^2 \leq 2\pi R \sum_{n=-\infty}^{\infty} |n|\,|f_n|^2 \leq R\|\nabla u\|^2_{L^2(\Omega)} \tag{C.52}$$

bounded, where the equality is (C.2). □

Lemma C.5. *The map $H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$ defined by*

$$u \times v \mapsto \int_{\partial\Omega} uB(k)v\,dS \tag{C.53}$$

is continuous.

Proof. Since the operator in (C.53) is bilinear, it is enough to show it is bounded. By density of $C^\infty(\bar{\Omega})$ in $H^1(\Omega)$, it is enough to show boundedness on $C^\infty(\bar{\Omega}) \times C^\infty(\bar{\Omega})$.

Take $u, v \in C^\infty(\bar{\Omega})$. Then $Tu \in L^2(\partial\Omega)$ by Corollary C.1 and $B(k)v \in L^2(\partial\Omega)$ by Lemma C.3, so

$$\int_{\partial\Omega} uB(k)v\,dS = 2\pi R \sum_{n=-\infty}^{\infty} f_{-n}B(k)_n g_n \tag{C.54}$$

from the proof of Lemma C.1. By Lemma C.2, for any $\varepsilon > 0$ there is an N such that $|B(k)_n| < |n|/R + \varepsilon$ for all $|n| > N$. Let $\bar{B}_N = \max_{|n|\leq N} |B(k)_n|$. Then

$$\Big|\sum_{n=-\infty}^{\infty} f_{-n}B(k)_n g_n\Big| \tag{C.55}$$

$$\leq \bar{B}_N \sum_{|n|\leq N} |f_{-n}|\,|g_n| + \sum_{|n|>N} |f_{-n}|\left(\frac{|n|}{R} + \varepsilon\right)|g_n| \tag{C.56}$$

$$\leq (\bar{B}_N + \varepsilon) \sum_{n=-\infty}^{\infty} |f_{-n}|\,|g_n| + \frac{1}{R}\sum_{n=-\infty}^{\infty} |f_{-n}|\,|n|\,|g_n| \tag{C.57}$$

$$= (\bar{B}_N + \varepsilon)\|(f_n)\|^2_{\ell^2}\|(g_n)\|^2_{\ell^2} + \frac{1}{R}\|(n^{1/2}f_n)\|^2_{\ell^2}\|(n^{1/2}g_n)\|^2_{\ell^2} \quad \text{(Cauchy-Schwarz)} \tag{C.58}$$

$$\leq \frac{(\bar{B}_N + \varepsilon)}{(2\pi R)^2}\|f\|^2_{L^2(\partial\Omega)}\|g\|^2_{L^2(\partial\Omega)} \quad \text{(by (C.2))} \tag{C.59}$$

$$+ \frac{1}{(2\pi)^2 R}\|\nabla u\|^2_{L^2(\Omega)}\|\nabla v\|^2_{L^2(\Omega)} \quad \text{(by Lemma C.4)} \tag{C.60}$$

$$\leq \left(\frac{\bar{B}_N + \varepsilon}{(2\pi R)^2} + \frac{1}{(2\pi)^2 R}\right)\|u\|^2_{H^1(\Omega)}\|v\|^2_{H^1(\Omega)}. \tag{C.61}$$

We have now shown

$$\left| \int_{\partial\Omega} uB(k)v \, dS \right| \leq C(\varepsilon, R, k)\|u\|^2_{H^1(\Omega)}\|v\|^2_{H^1(\Omega)}, \tag{C.62}$$

for all $u, v \in C^\infty(\bar{\Omega})$ which gives the result. □

Finally we can prove that (C.1) holds for all $u, v \in H^1(\Omega)$.

Theorem C.1. *If $u, v \in H^1(\Omega)$, then*

$$\int_{\partial\Omega} uB(k)v \, dS \; = \; \int_{\partial\Omega} vB(k)u \, dS. \tag{C.63}$$

Proof. By Lemma C.5, it is enough that (C.63) hold for all $u, v \in C^\infty(\bar{\Omega})$, because $C^\infty(\bar{\Omega})$ is dense in $H^1(\Omega)$ [Eva98]. Since (C.63) indeed holds for all $u, v \in C^\infty(\bar{\Omega})$ by Lemma C.3, the proof is complete. □

APPENDIX D

DERIVATION OF REALISTIC QUANTUM TUNNELING MODEL PARAMETERS

Harbury & Porod [HP96] used $m^* = 0.361m_0$ as the reduced mass of the electron and $E_B = 2.5$ eV as the potential height for 142.6 Angstrom diameter circular corrals made of 48 iron adatoms on copper. Their model differs from ours in that they are using finite cylinders centered at adatoms, whereas we use piecewise constant axisymmetric potentials. Despite the difference, we will use the heights and widths of the cylinders as estimates for the height and width of of our potential.

The time-independent Schrödinger equation is

$$\left(-\frac{\hbar^2}{2m}\Delta + V - E\right)\psi = 0 \tag{D.1}$$

where

- V is the potential function with units of energy (such as eV as above)
- Δ has units of $[L]^{-2}$ (such as $1/m^2$)
- $\hbar = 1.05 \times 10^{-34}\ m^2 \cdot kg/s = 6.58 \times 10^{-16}$ eV $\cdot\ s$
- $m = 0.361m_0$, where $m_0 = 9.11 \times 10^{-31}\ kg$ for the electron mass
- E is an energy equal to $\frac{\hbar^2 k^2}{2m}$ in terms of our frequency parameter k
- ψ has units of $[L]^{-1}$ so that its L^2-norm can be a unitless probability density.

The unitless version of the Schrödinger equation that we prefer to use is

$$\left(-\Delta + \tilde{V} - k^2\right)\tilde{\psi} = 0.$$

If we simply multiply (D.1) by $2m/\hbar^2$ we obtain

$$\left(-\Delta + \frac{2mV}{\hbar^2} - \frac{2mE}{\hbar^2}\right)\psi = 0.$$

Using the values for E_B and m^* above, we get

$$\frac{E_B}{\hbar} = \frac{2.5 \text{ eV}}{6.58 \times 10^{-16} \text{ eV} \cdot s} \approx 3.80 \times 10^{15}\ s^{-1},$$

$$\frac{m_0}{\hbar} = \frac{9.11 \times 10^{-31}\ kg}{1.05 \times 10^{-34}\ m^2 \cdot kg/s} \approx 8.68 \times 10^{3}\ s/m^2,$$

so

$$\frac{2m^* E_B}{\hbar^2} \approx 2 \cdot 0.361 \cdot 8.68 \times 10^{3}\ s/m^2 \cdot 3.80 \times 10^{15}\ s^{-1} \approx 2.38 \times 10^{19}\ 1/m^2$$

would be the height of our piecewise constant axisymmetric potential $\tilde{V}$ in units of $1/m^2$.

We could convert to Angstroms using 1 Angstrom $= 10^{-10}\ m$ and then put the radius of our corral as 71.3. If instead we define our own unit u so that the radius of this corral is $1\,u$, then the relationship between u and m is given by $1\,u = 71.3$ Angstroms $= 71.3 \times 10^{-10}\,m$. Then in terms of u, the height of $\tilde{V}$ is

$$\frac{(71.3)^2 \times 10^{-20}\ m^2}{1\ u^2} \cdot \frac{2m^* E_B}{\hbar^2} \approx 1204\ 1/u^2.$$

Now for the width of $\tilde{V}$. In [HP96], Harbury & Porod take the cylinders to be 1.52 Angstroms in diameter. So, proportionally, we want (potential width/corral diameter) = 1.52/142.6 which is about 0.0107. Therefore if our corral radius is 1 u, we want the potential width to be $0.0107 * 2\ u = 0.0214\ u$.

BIBLIOGRAPHY

[ACL09] A. Amiraslani, R. M. Corless, and P. Lancaster. Linearization of matrix polynomials expressed in polynomial bases. *IMA J. Numer. Anal.*, 29(1):141–157, 2009.

[AM11] Sk. Safique Ahmad and Volker Mehrmann. Perturbation analysis for complex symmetric, skew symmetric, even and odd matrix polynomials. *Electronic Transactions on Numerical Analysis*, 38:275–302, 2011.

[AR10] Fatima Aboud and Didier Robert. Asymptotic expansion for nonlinear eigenvalue problems. *Journal de Mathématiques Pures et Appliquées*, 93(2):149–162, Feb 2010.

[AS70] Milton Abramowitz and Irene Ann Stegun. *Handbook of mathematical functions : with formulas, graphs, and mathematical tables*. Dover Publications, New York, 9th edition, 1970.

[AST+09] Junko Asakura, Tetsuya Sakurai, Hiroto Tadano, Tsutomu Ikegami, and Kinji Kimura. A numerical method for nonlinear eigenvalue problems using contour integrals. *JSIAM Letters*, 1:52–55, 2009.

[AW05] George B. Arfken and Hans J. Weber. *Mathematical Methods for Physicists*. Elsevier Academic Press, 6th edition, 2005.

[BCG09] G.I. Bischi, C. Chiarella, and L. Gardini. *Nonlinear Dynamics in Economics, Finance and the Social Sciences: Essays in Honour of John Barkley Rosser Jr*. Springer Berlin Heidelberg, 2009.

[BEK11] Wolf-Jürgen Beyn, Cedric Effenberger, and Daniel Kressner. Continuation of eigenvalues and invariant pairs for parameterized nonlinear eigenvalue problems. *Numerische Mathematik*, 119(3):489, 2011.

[Ben02] Ivar Bendixson. Sur les racines d'une équation fondamentale. *Acta Mathematica*, 25:359–365, February 1902.

[Bey12] Wolf-Jürgen Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra and its Applications*, 436(10):3839–3863, 2012.

[BF60] F. L. Bauer and C. T. Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960.

[BF00] Dario Andrea Bini and Giuseppe Fiorentino. Design, analysis, and implementation of a multiprecision polynomial rootfinder. *Numerical Algorithms*, 23(2):127–173, 2000.

[BH13] David Bindel and Amanda Hood. Localization theorems for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1728–1749, 2013.

[BHM+13] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, and F. Tisseur. NLEVP: A collection of nonlinear eigenvalue problems. *ACM Trans. Math. Softw.*, 39(2):7:1–7:28, February 2013.

[BM00] A. Bellen and S. Maset. Numerical solution of constant coefficient linear delay differential equations as abstract Cauchy problems. *Numerische Mathematik*, 84(3):351–374, 2000.

[BMM13] Roel Van Beeumen, Karl Meerbergen, and Wim Michiels. A rational Krylov method based on Hermite interpolation for nonlinear eigenvalue problems. *SIAM Journal on Scientific Computing*, 35(1):A327–A350, 2013.

[BNS13] Dario A. Bini, Vanni Noferini, and Meisam Sharify. Locating the eigenvalues of matrix polynomials. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1708–1727, 2013.

[Bor10] Shreemayee Bora. Structured eigenvalue condition number and backward error of a class of polynomial eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 31(3):900–917, 2010.

[BZH10] Matthew C. Barr, Michael P. Zaletel, and Eric J. Heller. Quantum corral resonance widths: Lossy scattering as acoustics. *Nano Letters*, 10(9):3253–3260, 2010. PMID: 20684508.

[Car91] C. Carstensen. Inclusion of the roots of a polynomial based on Gerschgorin's theorem. *Numer. Math.*, 59(1):349–360, December 1991.

[CB96] S. Crampin and O. R. Bryant. Fully three-dimensional scattering

calculations of standing electron waves in quantum nanostructures: The importance of quasiparticle interactions. *Phys. Rev. B*, 54:R17367–R17370, Dec 1996.

[CD02] Younes Chahlaoui and Paul Van Dooren. Benchmark examples for model reduction of linear time invariant dynamical systems. MIMS EPrint 2008.22, Manchester Institute for Mathematical Sciences, The University of Manchester, UK., 2002.

[CGH+96] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth. On the Lambert W function. *Adv. Comput. Math.*, 5(1):329–359, 1996.

[Chu03] Eric King-wah Chu. Perturbation of eigenvalues for matrix polynomials via the Bauer–Fike theorems. *SIAM Journal on Matrix Analysis and Applications*, 25(2):551–573, 2003.

[CLE93] M. F. Crommie, C. P. Lutz, and D. M. Eigler. Confinement of electrons to quantum corrals on a metal surface. *Science*, 262(5131):218–220, 1993.

[CR00] Rebecca V. Culshaw and Shigui Ruan. A delay-differential equation model of HIV infection of CD4+ T-cells. *Mathematical Biosciences*, 165:27–39, 2000.

[CR01] J. Cullum and A. Ruehli. Pseudospectra analysis, nonlinear eigenvalue problems, and studying linear systems with delays. *BIT*, 41:265–281, 2001.

[DHT14] T. A. Driscoll, N. Hale, and L. N. Trefethen, editors. *Chebfun Guide*. Pafnuty Publications, Oxford, 2014.

[DSB92] W. Draijer, M. Steinbuch, and O.H. Bosgra. Adaptive control of the radial servo system of a compact disc player. *Automatica*, 28(3):455 – 462, 1992.

[DT03] Jean-Pierre Dedieu and Françoise Tisseur. Perturbation theory for homogeneous polynomial eigenvalue problems. *Linear Algebra Appl.*, 358(1–3):71 – 94, 2003.

[DTOBG08] Sergio Díaz-Tendero, Fredrik E. Olsson, Andrey G. Borisov, and

Jean-Pierre Gauyacq. Theoretical study of electron confinement in Cu corrals on a Cu(111) surface. *Phys. Rev. B*, 77:205403, May 2008.

[DZ16] Semyon Dyatlov and Maciej Zworski. Mathematical theory of scattering resonances, September 2016. `http://math.mit.edu/~dyatlov/res/res_20160915.pdf`.

[Eff13] C. Effenberger. Robust successive computation of eigenpairs for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1231–1256, 2013.

[EK12] C. Effenberger and D. Kressner. Chebyshev interpolation for nonlinear eigenvalue problems. *BIT*, 52:933–951, 2012.

[Els73] Ludwig Elsner. A remark on simultaneous inclusions of the zeros of a polynomial by Gershgorin's theorem. *Numerische Mathematik*, 21:425–427, 1973.

[ET01] Mark Embree and Lloyd N. Trefethen. Generalizing eigenvalue theorems to pseudospectra theorems. *SIAM Journal on Scientific Computing*, 23(2):583–590, 2001.

[Eva98] Lawrence Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I, 1998.

[FH03] Gregory A. Fiete and Eric J. Heller. *Colloquium* : Theory of quantum corrals and quantum mirages. *Rev. Mod. Phys.*, 75:933–948, Jul 2003.

[For] Steven Fortune. Eigensolve. `http://ect.bell-labs.com/who/sjf/eigensolve.html`. Accessed: 2 Nov 2016.

[For02] Steven Fortune. An iterated eigenvalue algorithm for approximating roots of univariate polynomials. *Journal of Symbolic Computation*, 33(5):627–646, May 2002.

[FS68] Avner Friedman and Marvin Shinbrot. Nonlinear eigenvalue problems. *Acta Mathematica*, 121(1):77, 1968.

[FT01] Karl Meerbergen Françoise Tisseur. The quadratic eigenvalue problem. *SIAM Review*, 43(2):235–286, 2001.

[Gal16] Jeffrey Galkowski. Resonances for thin barriers on the circle. *Journal of Physics A: Mathematical and Theoretical*, 49(12):125205, 2016.

[GDTB09] J.P. Gauyacq, S. Díaz-Tendero, and A.G. Borisov. Mapping of the electron transmission through the wall of a quantum corral. *Surface Science*, 603(13):2074 – 2081, 2009.

[Ger31] S. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS. Class des sciences mathématiques et naturelles.*, 6:749–754, 1931.

[GG94] Thomas Gebhardt and Siegfried Grossmann. Chaos transition despite linear stability. *Phys. Rev. E*, 50:3705–3711, Nov 1994.

[GGK90] Israel Gohberg, Seymour Goldberg, and Marinus A. Kaashoek. *Classes of Linear Operators Vol. I*. Birkhäuser Basel, 1990.

[GHA94] George R. Gray, David Huang, and Govind P. Agrawal. Chaotic dynamics of semiconductor lasers with phase-conjugate feedback. *Phys. Rev. A*, 49:2096–2105, Mar 1994.

[Giv99] D. Givoli. Recent advances in the DtN FE Method. *Archives of Computational Methods in Engineering*, 6(2):71–116, 1999.

[GK02] Kirk Green and Bernd Krauskopf. Global bifurcations and bistability at the locking boundaries of a semiconductor laser with phase-conjugate feedback. *Phys. Rev. E*, 66:016220, Jul 2002.

[GL96] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

[GLR09] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Society for Industrial and Applied Mathematics, 2009.

[Gov00] W. J. F. Govaerts. *Numerical Methods for Bifurcations of Dynamical Equilibria*. SIAM, 2000.

[Gri05] David Griffiths. *Introduction to Quantum Mechanics*. Pearson Prentice Hall, Upper Saddle River, NJ, 2005.

[GW06] Kirk Green and Thomas Wagenknecht. Pseudospectra and delay differential equations. *Journal of Computational and Applied Mathematics*, 196(2):567 – 578, 2006.

[HB] Amanda Hood and David Bindel. Pseudospectral bounds on transient growth for higher order and constant delay differential equations. *SIAM Journal on Matrix Analysis and Applications*, submitted. `https://arxiv.org/abs/1611.05130`.

[HCLE94] E. J. Heller, M. F. Crommie, C. P. Lutz, and D. M. Eigler. Scattering and absorption of surface electron waves in quantum corrals. *Nature*, 369(6480):464–466, 06 1994.

[Hir02] M. A. Hirsch. Sur les racines d'une équation fondamentale. *Acta Mathematica*, 25(1):367–370, 1902.

[HL93] J.K. Hale and S.M.V. Lunel. *Introduction to Functional Differential Equations*, volume 99 of *Applied Mathematical Sciences*. Springer, 1993.

[HLT07] N. J. Higham, R.-C. Li, and F. Tisseur. Backward error of polynomial eigenproblems solved by linearization. *SIAM J. Matrix Anal. Appl.*, 29:1218–1241, 2007.

[Hou64] Alton S. Householder. *The Theory of Matrices in Numerical Analysis*. Blaisdell Publishing Company, 1964.

[HP96] Henry K. Harbury and Wolfgang Porod. Elastic scattering theory for electronic waves in quantum corrals. *Phys. Rev. B*, 53:15455–15458, Jun 1996.

[HP06] D. Hinrichsen and A.J. Pritchard. *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*. Texts in Applied Mathematics. Springer, 2006.

[HT02] N. J. Higham and F. Tisseur. More on pseudospectra for polynomial eigenvalue problems and applications in control theory. *Linear Algebra Appl.*, 351–352(0):435–453, 2002.

[Jar08] E. Jarlebring. *The spectrum of delay-differential equations: numerical methods, stability and perturbation*. PhD thesis, Inst. Comp. Math, TU Braunschweig, 2008.

[JMM14] Elias Jarlebring, Karl Meerbergen, and Wim Michiels. Computing a partial Schur factorization of nonlinear eigenvalue problems using the infinite Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 35(2):411–436, 2014.

[Kau06] Linda Kaufman. Eigenvalue problems in fiber optic design. *SIAM Journal on Matrix Analysis and Applications*, 28(1):105–117, 2006.

[KGL98] Bernd Krauskopf, George R. Gray, and Daan Lenstra. Semiconductor laser with phase-conjugate feedback: Dynamics and bifurcations. *Phys. Rev. E*, 58:7190–7197, Dec 1998.

[Kim09] Seungil Kim. *Analysis of a PML Method Applied to Computation of Resonances in Open Systems and Acoustic Scattering Problems*. PhD thesis, Texas A & M University, College Station, TX, USA, 2009. AAI3384266.

[Kre09] Daniel Kressner. A block Newton method for nonlinear eigenvalue problems. *Numerische Mathematik*, 114(2):355–372, 2009.

[KS02] I. Kubiaczyk and S. H. Saker. Oscillation and stability in nonlinear delay differential equations of population dynamics. *Math. Comput. Model.*, 35(3-4):295–301, February 2002.

[LBLQL10] Ben-Shan Liao, Zhaojun Bai, and Kwok Ko Lie-Quan Lee. Nonlinear Rayleigh-Ritz iterative method for solving large scale nonlinear eigenvalue problems. *Taiwanese Journal of Mathematics*, 14(3A):869–883, June 2010.

[LBLV92] B. Lehman, J. Bentsman, S. V. Lunel, and E. I. Verriest. Vibrational control of nonlinear time lag systems with arbitrarily large but bounded delay: averaging theory, stabilizability, and transient behavior. In *Decision and Control, 1992., Proceedings of the 31st IEEE Conference on*, pages 1287–1294 vol.2, 1992.

[Leh94] B. Lehman. Stability of chemical reactions in a CSTR with delayed recycle stream. In *American Control Conference*, volume 3, pages 3521–3522, June 1994.

[Lév81] Lucien Lévy. Sur la possibilité de l'équilibre électrique. *Comptes rendus de l'Académie des Sciences*, XCIII:706–708, 1881.

[Lia07] Ben-Shan Liao. *Subspace Projection Methods for Model Order Reduction and Nonlinear Eigenvalue Computation*. PhD thesis, University of California, Davis, Davis, CA, USA, 2007. AAI3280614.

[Lin02] K. K. Lin. Numerical study of quantum resonances in chaotic scattering. *Journal of Computational Physics*, 176:295–329, March 2002.

[Mel13] A. Melman. Generalization and variations of Pellet's theorem for matrix polynomials. *Linear Algebra and its Applications*, 439:1550–1567, 2013.

[MGWN06] Wim Michiels, Kirk Green, Thomas Wagenknecht, and Silviu-Iulian Niculescu. Pseudospectra and stability radii for analytic matrix functions with application to time-delay systems. *Linear Algebra and its Applications*, 418(1):315 – 335, 2006.

[Min00] Hermann Minkowski. Zur Theorie der Einheiten in den algebraischen Zahlkörpern. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1900:90–93, 1900.

[MMMM06] D. Steven Mackey, Niloufer Mackey, Christian Mehl, and Volker Mehrmann. Vector spaces of linearizations for matrix polynomials. *SIAM Journal on Matrix Analysis and Applications*, 28(4):971–1004, 2006.

[MN07a] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics, 3rd edition, 2007.

[MN07b] Wim Michiels and Silviu-Iulian Niculescu. *Stability and Stabilization of Time-Delay Systems (Advances in Design & Control)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[Moi98] Nimrod Moiseyev. Quantum theory of resonances: calculating energies, widths and cross-sections by complex scaling. *Physics Reports*, 302(5–6):212 – 293, 1998.

[Mos86] Ronald G Mosier. Root neighborhoods of a polynomial. *Math. Comput.*, 47(175):265–273, July 1986.

[MV04] Volker Mehrmann and Heinrich Voss. Nonlinear eigenvalue problems: A challenge for modern eigenvalue methods. *GAMM Mitteilungen*, 27:121–152, 2004.

[MW02] Volker Mehrmann and David Watkins. Polynomial eigenvalue problems with Hamiltonian structure. *Electronic Transactions on Numerical Analysis*, 13:106–118, 2002.

[Pen87] Louis L. Pennisi. Coefficients of the characteristic polynomial. *Mathematics Magazine*, 60(1):31–33, February 1987.

[Pli05] Elmar Plischke. *Transient Effects of Linear Dynamical Systems*. PhD thesis, Universität Bremen, July 2005.

[RI11a] Rizwana Rehman and Ilse C. F. Ipsen. Computing characteristic polynomials from eigenvalues. *SIAM Journal on Matrix Analysis and Applications*, 32(1):90–114, 2011.

[RI11b] Rizwana Rehman and Ilse C. F. Ipsen. La Budde's method for computing characteristic polynomials, 2011. `https://arxiv.org/abs/1104.3769`.

[Rou62] Eugène Rouché. Mémoire sur la série de Lagrange. *Journal de l'école impériale polytechnique*, 22(39):193–224, 1862.

[Rud76] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill, New York, 1976.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill, New York, 1987.

[RZ04] A. I. Rahachou and I. V. Zozoulenko. Elastic scattering of surface electron waves in quantum corrals: Importance of the shape of the adatom potential. *Phys. Rev. B*, 70:233409, Dec 2004.

[SB11] Yangfeng Su and Zhaojun Bai. Solving rational eigenvalue problems via linearization. *SIAM Journal on Matrix Analysis and Applications*, 32(1):201–216, 2011.

[Sch08] Kathrin Schreiber. *Nonlinear eigenvalue problems: Newton-type methods and nonlinear Rayleigh functionals*. PhD thesis, Technische Universität Berlin, 2008.

[Sin08] John R. Singler. Transition to turbulence, small disturbances, and sensitivity analysis I: A motivating problem. *Journal of Mathematical Analysis and Applications*, 337(2):1425 – 1441, 2008.

[Sol06] Sergey I. Solov'ëv. Preconditioned iterative methods for a class of nonlinear eigenvalue problems. *Linear Algebra and its Applications*, 415(1):210–229, May 2006.

[SS03] E.B. Saff and A.D. Snider. *Fundamentals of Complex Analysis with Applications to Engineering and Science*. Prentice Hall, 2003.

[TE05] L.N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

[TH01] Françoise Tisseur and Nicholas J. Higham. Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM Journal on Matrix Analysis and Applications*, 23(1):187–208, 2001.

[Tis00] Françoise Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309(1–3):339 – 361, 2000.

[Tre00] L. Trefethen. *Spectral Methods in MATLAB*. Society for Industrial and Applied Mathematics, 2000.

[Var04] R. Varga. *Gerschgorin and his circles*. Springer, 2004.

[VBJM16] Roel Van Beeumen, Elias Jarlebring, and Wim Michiels. A rank-exploiting infinite Arnoldi algorithm for nonlinear eigenvalue problems. *Numerical Linear Algebra with Applications*, 23(4):607–628, 2016. nla.2043.

[VBK16] Marc Van Barel and Peter Kravanja. Nonlinear eigenvalue problems and contour integrals. *J. Comput. Appl. Math.*, 292(C):526–540, January 2016.

[Vos13] Heinrich Voss. Nonlinear eigenvalue problems. In Leslie Hogben, editor, *Handbook of Linear Algebra, Second Edition*, chapter 60. Chapman and Hall/CRC, 2013.

[Wei12] H.F. Weinberger. *A First Course in Partial Differential Equations: with Complex Variables and Transform Methods*. Dover Books on Mathematics. Dover Publications, 2012.

[Wlo87] J. Wloka. *Partial Differential Equations*. Cambridge University Press, Cambridge, 1987.

[WMG08] T. Wagenknecht, W. Michiels, and K. Green. Structured pseudospectra for nonlinear eigenvalue problems. *J. Comput. Appl. Math.*, 212(2):245–259, February 2008.

[XH10] Yangfeng Su Xin Huang, Zhaojun Bai. Nonlinear rank-one modification of the symmetric eigenvalue problem. *Journal of Computational Mathematics*, 28(2):218–234, 2010.

[XMZZ16] Jinyou Xiao, Shuangshuang Meng, Chuanzeng Zhang, and Changjun Zheng. Resolvent sampling based Rayleigh–Ritz method for large-scale nonlinear eigenvalue problems. *Computer Methods in Applied Mechanics and Engineering*, 310:33 – 57, 2016.

[Zha05] Fuzhen Zhang. *The Schur complement and its applications*. Springer, New York, 2005.